Université de Montréal

Inférence bayésienne pour la reconstruction d'arbres phylogénétiques

par

Javier Oyarzun

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

juin 2006

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Inférence bayésienne pour la reconstruction d'arbres phylogénétiques

présenté par

Javier Oyarzun

a été évalué par un jury composé des personnes suivantes :

Yves Lepage
(président-rapporteur)

Jean-François Angers
(directeur de recherche)

Bernard Angers
(co-directeur)

Pierre Duchesne
(membre du jury)

Mémoire accepté le:
Le 21 juin 2006

SOMMAIRE

Ce mémoire a pour but de démontrer l'utilisation d'une approche bayésienne dans la reconstruction d'arbres phylogénétiques. Ainsi, nous voulons déterminer la légitimité de prendre une nouvelle approche dans un domaine où la méthode du maximum de vraisemblance (Edwards et Cavalli-Sforza, 1963) est considérée comme la norme. Pour ce faire, nous expliquons l'histoire de la phylogénétique et l'évolution des méthodes de reconstruction d'arbres phylogénétiques. Ensuite, nous illustrons nos méthodes informatiques de représentation d'arbres et de calcul de la vraisemblance des arbres phylogénétiques. Tout en démontrant les subtilités des modèles de substitution des données moléculaires, nous élaborons les propriétés de l'approche bayésienne et l'utilisation de la procédure Metropolis-Hastings (Metropolis *et al.*, 1953 ; Hastings, 1970) pour pouvoir effectuer l'algorithme du MCMC.

Afin de tester la validité d'une telle approche, nous développons des tests de comparaison pour pouvoir mesurer la performance de l'approche bayésienne face à la méthode du maximum de vraisemblance. Finalement, nous terminons en comparant les méthodes à l'aide d'échantillons de jeux de données simulées et à partir d'échantillons de jeux de données réelles. Cette comparaison permet de constater les nombreux avantages que possède l'approche bayésienne vis-à-vis la méthode du maximum de vraisemblance.

SUMMARY

The aim of this master thesis is to show the use of a Bayesian approach in the building of phylogenetic trees. Thus, we want to determine the legitimacy of taking this new approach in a field where the method of the maximum of likelihood (Edwards and Cavalli-Sforza, 1963) is regarded as the gold standard. With this intention, we explain the history of phylogenetic and the evolution of its methods.

Afterward, we illustrate our data-processing methods of representation of trees and calculation of the likelihood of the phylogenetic trees. While showing subtleties of the models of substitution of the molecular data, we work out the properties of the Bayesian approach and the use of the Metropolis-Hastings procedure (Metropolis *et al.*, 1953 ; Hastings, 1970) to carry out the MCMC algorithm.

In order to test the validity of such an approach, we used statistical tests to compare our approach against the maximum likelihood method. Finally, we finish by comparing the methods using samples of simulated data and samples of real data. This comparison makes it possible to note the many advantages of a Bayesian approach in the field of phylogenetics.

MOTS CLÉS

Phylogénie, méthode de vraisemblance, représentation graphique d'arbres, modèles de substitution, Jukes-Cantor, algorithme EM, catalogage d'arbres, distance entre arbres.

KEYWORDS

Phylogeny, likelihood method, graphical representation of trees, substitution models, Jukes-Cantor, EM algorithm, cataloguing of trees, distance between trees.

REMERCIEMENTS

En premier lieu, je tiens à remercier mon directeur de recherche, M. Jean-François Angers, qui a su faire preuve d'une patience exemplaire. Son soutien remarquable, ainsi que son dévouement furent d'un grand secours dans l'aboutissement de cette recherche.

En second lieu, je me dois remercier mon co-directeur de recherche, M. Bernard Angers, qui a eu la gentillesse de répondre à toutes mes questions, sans jamais perdre patience.

J'aimerais également exprimer ma reconnaissance envers le Département de mathématiques et de statistique de l'Université de Montréal pour leur support financier.

Je tiens à remercier mes collègues de bureau, Louis, Marie-Ève et Vincent qui eurent à m'endurer et à m'encourager lors de l'écriture de ce mémoire. Le parcours n'aurait pu être aussi agréable sans leur présence et celle de tous mes amis, professeurs et employés du département.

Finalement, je dédie ce présent mémoire à mes parents, à mon frère et à mon directeur.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

La compréhension de soi passe souvent par la connaissance de nos ancêtres. Comme la majorité des êtres humains, nous cherchons à en savoir plus sur notre généalogie. Toutefois, celle-ci peut aller aussi loin que la première cellule vivante. Une telle reconstruction porte l'appellation phylogénie. Les phylogénies sont à la base de la compréhension des différences entre espèces. Le terme phylogénie fut inventé par Ernst Haeckel en 1866 pour définir l'enchaînement des espèces vivantes au cours du temps.

Depuis 140 ans, les chercheurs développent des nouvelles méthodes pour différencier les espèces, afin de pouvoir reconstruire l'arbre de l'évolution. Les premières phylogénies publiées par Haeckel (1866) intégrant divers mammifères actuels et fossiles furent construites à partir des différences et similarités des traits morphologiques des espèces étudiées. Jusqu'à l'oeuvre de Hennig (1950), la construction phylogénétique obéissait au principe du triple parallélisme qui soutient que l'anatomie comparée, l'ontogénie et la paléontologie fournissent les sources de la reconstruction phylogénétique. En regroupant les espèces et cherchant à identifier les états primitifs des traits morphologiques, Hennig introduit une approche plus systématique, dite cladistique, pour comparer les espèces. Indépendamment de la systématique phylogénétique de Hennig, Edwards et Cavalli-Sforza (1963) ont invoqué explicitement le principe de la parcimonie qui désigne l'estimation de l'arbre phylogénétique comme étant l'arbre qui fait appel à la quantité minimale d'évolution (minimum de mutations).

Jusqu'aux années 1960, les seuls caractères descriptifs disponibles étaient les traits morphologiques, les comportements et la répartition géographiques des espèces. Les progrès de la biologie moléculaire ont produit des recherches sur les

acides nucléiques, et par ce fait même les acides aminés, qui ont permis d'étudier les espèces à partir de leurs gènes. Ces recherches amenèrent les chercheurs à développer de nouvelles méthodes statistiques, telles que les méthodes d'estimation de maximum de vraisemblance en phylogénétique. Les méthodes par maximum de vraisemblance furent introduites par Edwards et Cavalli-Sforza (1964) et Neyman (1971). Felsenstein (1981) développa les premiers algorithmes informatiques pour calculer les vraisemblances dans le domaine de la phylogénétique. Depuis, des méthodes de maximum de vraisemblance avec des modèles évolutifs plus complexes furent développés tels que ceux de Kimura (1981), Lanave *et al.* (1984), Tavaré (1986), Tamura et Nei (1993) et Zharkikh (1994). En parallèle, plusieurs autres méthodes de reconstruction d'arbres phylogénétiques furent développées telles que les méthodes de distances (Edwards et Cavalli-Sforza, 1963 ; Fitch et Margoliash, 1967), certaines variantes de la méthode de parcimonie (Camin-Sokal, 1965 ; Farris, 1970), lesquelles ne seront pas décrites en détail dans ce mémoire.

Une approche bayésienne dans la reconstruction d'arbres phylogénétiques fut introduite par Yang et Rannala (1997), Mau et Newton (1997), Larget et Simon (1999) et finalement Li, Pearl et Doss (2000). Cette approche se base sur le théorème de Bayes (1763) et de l'utilisation de la méthode de Monte Carlo par chaînes de Markov (Metropolis *et al.*, 1953) pour reconstruire des arbres phylogénétiques. Cette méthode qui est à son début fait l'oeuvre de plusieurs recherches dont celle de Huelsenbeck *et al.* (2001) qui ont développé le programme MrBayes, le programme de reconstruction d'arbres phylogénétiques employant une approche bayésienne le plus utilisé.

Dans ce mémoire, nous optons plutôt pour une modélisation bayésienne dans la reconstruction d'arbres phylogénétiques à partir de données moléculaires, par opposition à des jeux de données morphologiques, géographiques ou autres. Ensuite, nous proposons différents tests permettant de comparer notre approche bayésienne aux méthodes de maximum de vraisemblance. L'estimation des arbres se fera à partir d'échantillons de jeux de données simulées et à partir d'échantillons de vrais jeux de données. Conséquemment, les tests permettront de valider l'approche bayésienne pour une grande variété d'arbres phylogénétiques.

En résumé, ce mémoire est organisé de la manière suivante. Dans le premier chapitre, nous nous intéressons à expliquer les fondements de la phylogénétique. Nous présentons une introduction aux termes biologiques qui sont utilisés dans le mémoire. Nous illustrons les concepts de la génétique et de la théorie de l'évolution qui régissent la phylogénétique moléculaire. Ensuite, nous abordons le concept de la théorie des graphes pour expliquer le concept de la topologie des arbres phylogénétiques. Finalement, nous introduisons les manières d'énumérer les arbres et nous élaborerons brièvement les méthodes de reconstruction des arbres phylogénétiques.

Dans le deuxième chapitre, nous présentons nos méthodes pour rallier la phylogénétique et l'informatique. Dans un premier temps, nous expliquons notre représentation informatique d'un arbre phylogénétique. En deuxième temps, nous montrons notre méthode pour cataloguer les arbres. Finalement, nous décrivons notre procédé informatique de calcul de vraisemblance pour un arbre phylogénétique.

Au troisième chapitre, nous nous intéressons à expliquer les fonctionnements des divers modèles de substitution qui permettent de donner une vraisemblance aux arbres phylogénétiques. De plus, ce chapitre démontre les manières d'obtenir les modèles de substitution. En dernier lieu, nous décrivons nos méthodes d'optimisation de longueurs de branches en utilisant l'algorithme EM (Dempster, Laird et Rubin, 1977).

Dans le quatrième chapitre, nous élaborons nos procédés pour utiliser une approche bayésienne dans la reconstruction d'arbres phylogénétiques. Nous expliquons la base de l'approche, du théorème de Bayes (1763) à l'utilisation de la méthode de Monte Carlo par chaînes de Markov (Metropolis *et al.*, 1953) et la méthode Metropolis-Hastings (Metropolis *et al.*, 1953 ; Hastings, 1970). Nous présentons ensuite nos propositions de sélection d'arbres, ainsi que celles de Larget *et al.* (2005). Toutes les propriétés sur l'approche bayésienne sont expliquées en détail. Nous terminons ce chapitre en illustrant les méthodes de comparaison qui permettent de valider l'approche bayésienne dans le domaine de la phylogénétique.

Finalement, au dernier chapitre, nous étudions le comportement de notre approche bayésienne à l'aide de simulations, puis nous comparons les résultats obtenus à ceux présentés par les méthodes de maximum de vraisemblance. Pour ce faire, nous utilisons des jeux de données simulées et des jeux de données réelles.

Chapitre 1

INTRODUCTION À LA PHYLOGÉNÉTIQUE

1.1. Historique de la phylogénétique

L'histoire de la phylogénétique commence avec les découvertes de Darwin. En fait, Darwin fut l'un des premiers à présenter un arbre généalogique, dit phylogénétique, de l'évolution du vivant. Ainsi, si nous descendons tous les uns des autres, il est possible de construire un arbre phylogénétique des espèces vivantes. La phylogénétique est une approche de classification biologique qui cherche à regrouper les espèces par la comparaison de caractères homologues. Jusqu'aux années 1960, les seuls caractères descriptifs disponibles étaient les traits morphologiques (ex. présence d'ailes, présence de la corde dorsale, etc.), les comportements et la répartition géographiques des espèces. Ceci pouvait soulever des débats sur l'objectivité de la phylogénie, puisque le nombre de caractères était limité. Les progrès de la biologie moléculaire ont produit des recherches sur les acides nucléiques, et par ce fait même les acides aminés, qui ont permis d'étudier les espèces à partir de leurs gènes. En conséquence, les chercheurs disposent d'un plus grand nombre de caractères à comparer, et ceci donne une plus grande objectivité aux travaux en phylogénétique. Ainsi aujourd'hui, la phylogénie est presque entièrement liée aux recherches sur les acides nucléiques, prénommé la phylogénie moléculaire.

1.2. Les gènes

Les fondements de la génétique naquirent avec les travaux de Mendel sur le pois (1866) et de Thomas H. Morgan sur les mouches drosophiles (1910) (voir

l'ouvrage de Bateson, 1902). Ces chercheurs ont mis en évidence l'existence de facteurs biologiques de l'hérédité. La transmission de ces facteurs, dans le cas de caractères simples, pouvait s'expliquer par l'existence d'entités d'information génétique discrètes : les gènes. Un gène désigne une unité d'information génétique transmise par un individu à sa descendance, de façon sexuée ou asexuée. Le gène consiste en un segment d'acide désoxyribonucléique (ADN) traduit en acide ribonucléique (ARN), cet ARN est utilisé par la cellule pour transmettre l'information à l'extérieur du noyau, puis pour synthétiser des protéines à partir de ces informations. L'ensemble des gènes d'un individu constitue entre 3 % et 10 % de son génome, le reste étant de l'ADN non codant. Il existe environ 18 000 gènes dans l'ADN des cellules d'une drosophile et 30 000 gènes dans celles d'un humain. Un gène est caractérisé par sa séquence de nucléotides (voir section 1.2.1). Un seul gène peut regrouper entre 15 et 10 000 acides nucléiques (nucléotides).

1.2.1. Acides nucléiques

Les nucléotides (acides nucléiques) sont des acides désoxyribonucléiques pour l'ADN et ribonucléiques pour l'ARN. Un nucléotide est composé de 3 parties :

(1) un groupement de phosphate ;

(2) un sucre à 5 atomes de carbone (désoxyribose pour l'ADN et ribose pour l'ARN) ;

(3) une base azotée variable en fonction du nucléotide.

Il existe quatre types de nucléotide pour l'ADN : l'adénine (noté A), la cytosine (noté C), la guanine (noté G) et la thymine (noté T) (voir figure 1.2.1). Ces nucléotides ont la particularité de s'unir deux à deux par complémentarité. L'adénine se complémente avec la thymine et la cytosine avec la guanine. Les bases de chaque brin d'ADN s'apparient par des liaisons hydrogènes : deux liaisons hydrogènes pour les appariements de base A-T (adénine-thymine) et trois liaisons pour G-C (cytosine-guanine). Les nucléotides de l'ARN sont les mêmes que pour l'ADN sauf la thymine (T) qui est remplacée par l'uracile (noté U) qui s'apparie avec l'adénine (A).

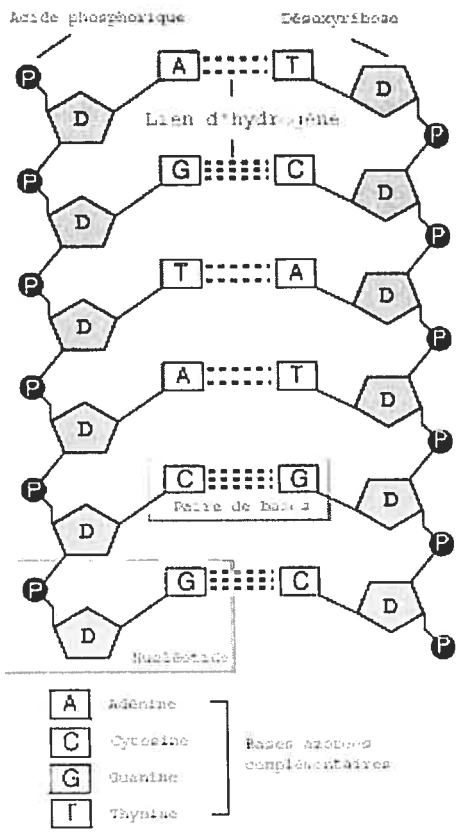


FIGURE 1.2.1. Acide nucléique (nucléotide)

1.2.2. Gène mitochondrial

Le matériel génétique (ADN mitochondrial) de la mitochondrie (qui est la seule partie des cellules animales à posséder son propre ADN, en plus du noyau) n'a subi que très peu de modifications depuis le début de l'évolution et sert souvent dans les recherches phylogénétiques. Le génome mitochondrial est composé de 15 000 à 21 000 paires de bases azotées.

Chez les animaux, lors de la reproduction, les mitochondries du spermatozoïde ne peuvent pas (en général) se transmettre dans l'ovocyte. Par conséquent, toutes les mitochondries d'un individu lui sont transmises par sa mère. L'étude de l'ADN mitochondrial permet donc de retracer les relations généalogiques entre les individus selon la voie maternelle. Ainsi, tous les animaux possèdent ces mêmes

gènes mitochondriaux, il devient donc possible de construire des généalogies pour les espèces vivantes (Campbell, 1995).

1.2.3. Mutation

Une mutation représente le fait de changer, d'être modifié. En biologie, le terme mutation est utilisé pour illustrer le changement de la séquence des bases, séquence dont l'ordonnancement donne toute sa signification à l'ADN et lui permet de coder les gènes. Donc, une mutation peut être représentée par un changement d'un nucléotide de thymine (T) pour un d'adénine (A) (voir figure 1.2.2).

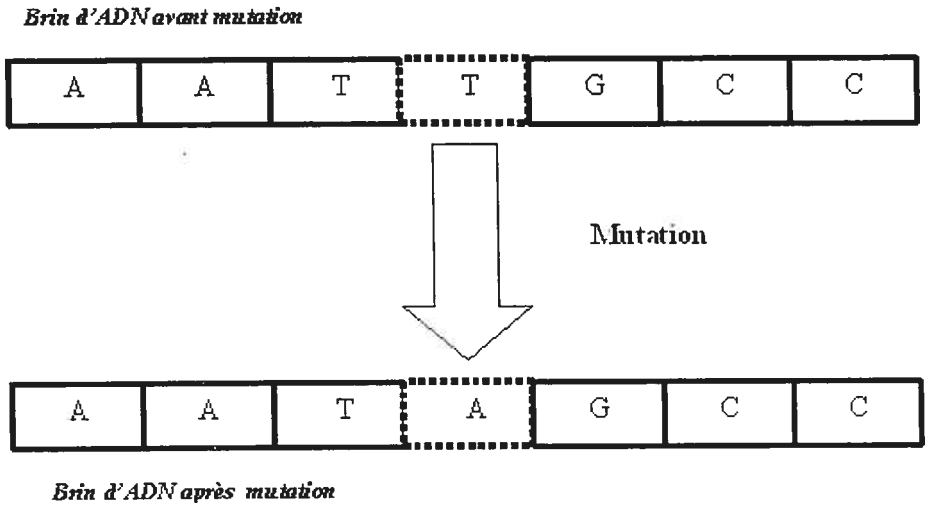


FIGURE 1.2.2. Mutation d'un nucléotide

1.2.3.1. *Transitions et transversions*

Les mutations se subdivisent en deux catégories : les transitions et les transversions. Les transitions sont les mutations (substitutions) qui mènent une base (nucléotide) A à une base G ou vice-versa d'une base G à une base A. Les nucléotides A et G appartiennent au groupe des purines. Il existe également les transitions qui transforment une base C à une base T ou vice-versa d'une base T à une base C. Les nucléotides C et T appartiennent au groupe des pyrimidines. Donc, une transition est une mutation de purine à purine ou de pyrimidine à

pyrimidine. En ce qui concerne les transversions, elles sont composées de toutes les autres mutations possibles : de A à C, de C à A, de A à T, de T à A, de C à G, de G à C, de G à T et de T à G. En conséquence, les transversions sont des mutations qui transforment une purine en une pyrimidine ou qui transforment une pyrimidine en une purine (voir figure 1.2.3).

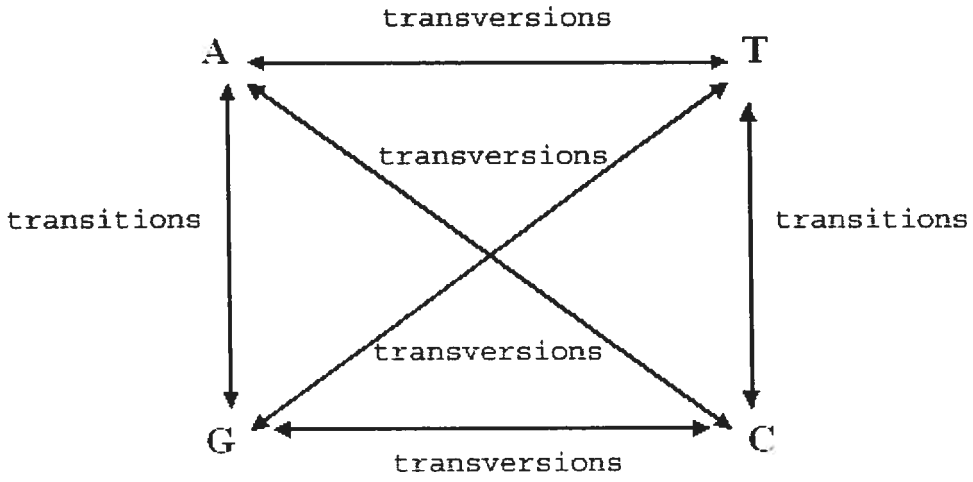


FIGURE 1.2.3. Transversions et transitions

1.2.4. La théorie de l'évolution

Une observation rigoureuse des êtres vivants montre l'existence de nombreux points communs entre les espèces vivantes. La théorie de l'évolution explique ces ressemblances par l'existence de liens généalogiques entre toutes les formes de vie : les organismes se ressemblent parce qu'ils partagent des caractères hérités d'un ancêtre commun. De nombreuses disciplines scientifiques apportent des preuves en faveur de l'évolution :

(1) les preuves anatomiques : les différentes espèces d'un embranchement, malgré des aspects extérieurs très différents, partagent un plan d'organisation invariable (tous les vertébrés ont une chaîne nerveuse dorsale et une tête vers l'avant du corps, par exemple).

(2) Les preuves biologiques : chez toutes les espèces, il existe une certaine variabilité. L'éventail de variations que présente une population vivante est le matériel de base avec lequel l'évolution peut construire des organismes de plus en plus différents.

(3) Les preuves génétiques : tous les êtres vivants fonctionnent sur les mêmes bases moléculaires (ADN et ARN). Ils utilisent le même code génétique (à quelques exceptions près). Les différences entre les espèces et les variations entre individus au sein d'une même espèce sont dues à des différences dans la séquence des gènes et la structure des chromosomes, différences provoquées par des mutations et des réarrangements chromosomiques occasionnels. Les individus dont les gènes sont légèrement différents de ceux de leurs parents, suffisent pour introduire la variabilité, la source de l'évolution.

(4) Les preuves paléontologiques : les fossiles des êtres vivants disparus témoignent de l'évolution de la vie depuis son apparition (datée d'approximativement 3,5 milliards d'années). Nous connaissons de nombreuses formes de vies, aujourd'hui disparues, faisant le lien entre différents groupes d'une lignée évolutive. Par exemple l'archaeopteryx, l'ancêtre des oiseaux et qui provient des dinosaures.

1.3. PHYLOGÉNIE

La science de la phylogénie est l'étude de la formation et de l'évolution des organismes en vue d'établir leur parenté. La phylogenèse est le terme le plus utilisé pour décrire la généalogie d'une espèce, d'un groupe d'espèces mais également, à un niveau intraspécifique, la généalogie entre populations ou entre individus. Elle est représentée par un arbre, dit arbre phylogénétique. Il décrit les connexions entre les espèces et leur distance génétique, c'est-à-dire le nombre de mutations dans le matériel génétique qui les séparent.

Définition 1.3.1. *Un arbre phylogénétique est un graphe orienté acyclique (voir section 1.3.2) qui montre les relations de parentés entre les espèces ou d'autres entités supposées avoir un ancêtre commun. Chacun des noeuds de l'arbre représente l'ancêtre commun de ses descendants.*

En regardant la figure 1.3.1, il est possible d'affirmer que l'espèce hypothétique A est l'ancêtre commun du serpent et de l'espèce hypothétique B, qui elle-même est l'ancêtre de la baleine et de l'espèce hypothètique C, et finalement que l'homme et le chimpanzé sont des descendants de l'espèce hypothétique C.

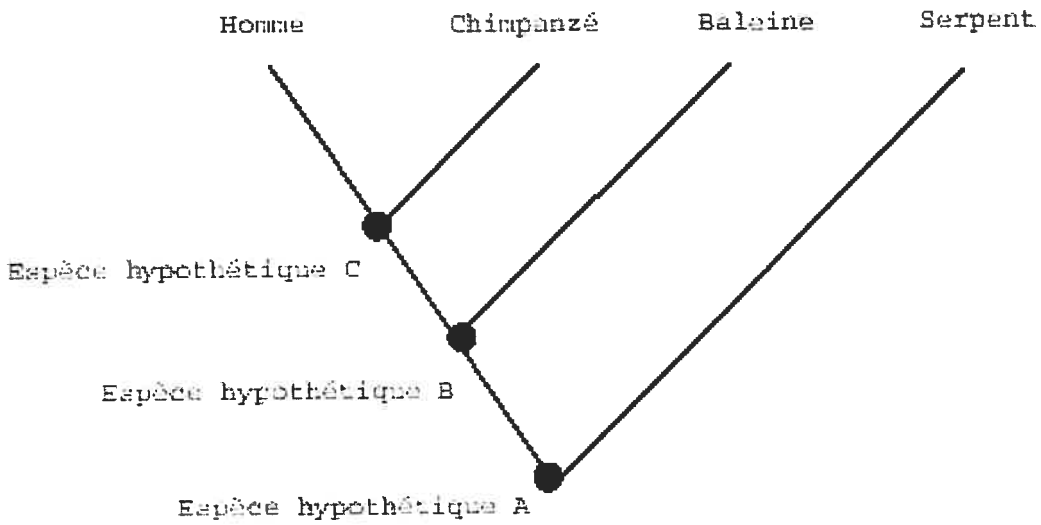


FIGURE 1.3.1. Un arbre phylogénétique

1.3.1. Les arbres

La méthode intuitive pour trouver l'arbre phylogénétique d'un groupe d'espèces, serait de construire tous les arbres possibles et de les comparer. L'arbre le plus probable, selon un modèle d'évolution donné, représenterait l'arbre phylogénétique de ces espèces. Toutefois, comme nous allons le démontrer dans cette section, il est souvent impossible de construire tous les arbres, puisqu'il existe

des multitudes de combinaisons possibles pour un regroupement d'espèces. Cette section s'efforcera d'expliquer la topologie des arbres.

1.3.2. Théorie des graphes

Définition 1.3.2. *Il existe deux types de graphes : les graphes orientés et les graphes non orientés. Dans les deux cas, un graphe G est un couple de deux ensembles (S,A) où S, a priori non vide, est appelé ensemble des sommets de G :*

(1) si G est un graphe orienté alors A est une partie de $S \times S$ (produit cartésien d'ensembles). Les éléments de A sont des couples de sommets (relation d'ordre) et se nomment arcs.

(2) Si G est un graphe non orienté alors A est une partie de $S \times S$. Les éléments de A sont des paires de sommets (il n'y a pas de relation d'ordre), et se nomment arêtes (droite reliant deux points).

(3) Deux sommets a et b d'un graphe orienté (resp. non orienté) $G = (S,A)$ sont dits reliés s'il existe un arc (a,b) ou (b,a) (respectivement une arête a,b) dans A.

Si un graphe possède deux sommets reliés par plusieurs arcs ou arêtes, nous tombons dans le cas de multigraphe dont la définition formelle est différente. Dans un graphe orienté (respectivement non orienté) une boucle est un arc (a, b) avec $a = b$ (resp. une arête à un seul sommet). Un graphe sans boucle est un graphe simple.

1.3.3. Arbre topologique

Un arbre est une représentation graphique d'une structure dont les espèces apparaissent sous la forme de noeuds qui, de haut en bas dominent des systèmes d'embranchement. Ainsi, l'arbre est composé des noeuds externes et internes (voir figure 1.3.2), qui représentent les espèces vivantes (en gris clair) et les espèces ancestrales (en gris foncé) respectivement, et de branches (en noir) qui lient les espèces les unes aux autres. Incidemment, il s'agit d'un graphe orienté où les sommets sont des noeuds et les arêtes sont des branches. Tout dépendamment du contexte, il est possible d'enraciner un arbre, ce qui signifie de lui fixer un ancêtre,

de cette manière nous obtenons un arbre qui évolue dans le temps (voir figure 1.3.2). Pour ce mémoire, l'enracinement de l'arbre, dit embranchement ancestral, sera utilisé pour construire les arbres.

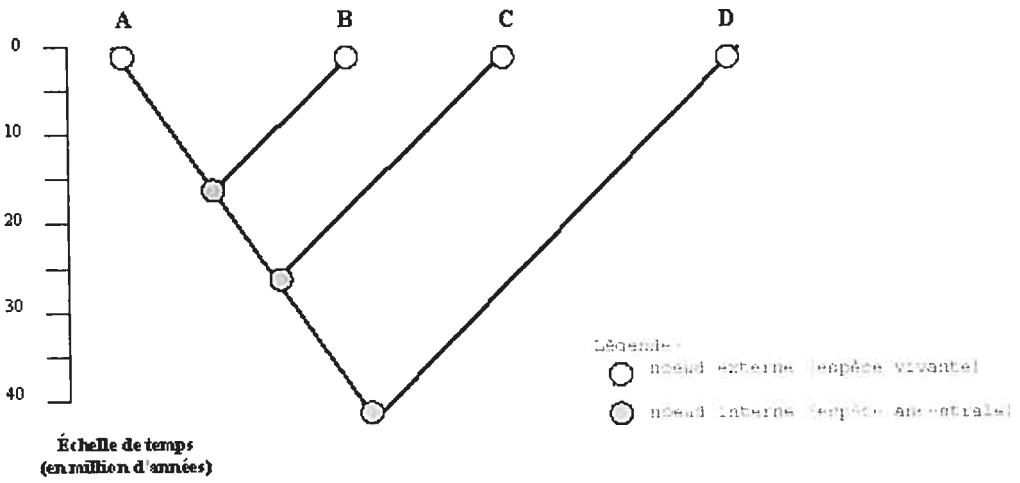


FIGURE 1.3.2. Les caractéristiques d'un arbre topologique (phylogénétique)

1.3.4. Construction des arbres

Il existe deux types d'arbre enraciné, les arbres dichotomiques et les arbres polyfurcaux. Les arbres dichotomiques sont les arbres dont seulement deux branches se joignent à un noeud. Tandis que pour les arbres polyfurcaux, les noeuds peuvent être embranchés par deux branches ou plus. La figure 1.3.3 montre l'ensemble des arbres dichotomiques possibles pour 2, 3 et 4 espèces. Ceci étant dit, la construction des arbres polyfurcaux se fait de la même manière, mais avec plus de nouveaux arbres à chaque fois que nous augmentons le nombre d'espèces. Les arbres polyfurcaux seront utilisés pour cette recherche, puisque les arbres dichotomiques sont un sous-ensemble des arbres polyfurcaux et qu'ils peuvent permettre de déceler des irrégularités dans les jeux de données. Les arbres polyfurcaux indiquent :

(1) une irrégularité dans un jeu de données, des espèces suivant des modèles d'évolution différents ;

(2) un phénomène de spéciation très rapide (changements géographiques).

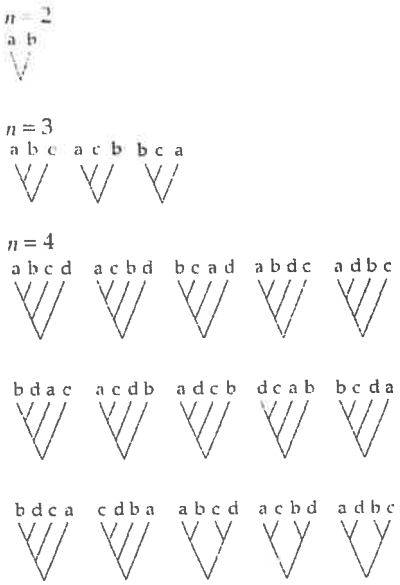


FIGURE 1.3.3. Les arbres dichotomiques de 2, 3 et 4 espèces (n)

TABLEAU 1.3.1. Énumération du nombre d'arbres en fonction du nombre de noeuds et d'espèces

	nombre de noeuds ancestraux					
		1	2	3	4	nombre total d'arbres
nombre	2	1				1
d'espèces	3	1	3			4
	4	1	10	15		26
	5	1	25	105	105	236

1.3.5. Énumération des arbres

En construisant tous les arbres polyfurcaux pour les phylogénies allant de 2 à 5 espèces, nous pouvons remarquer qu'il existe une relation récursive entre le nombre d'arbres et le nombre d'espèces dans la phylogénie.

La figure 1.3.4 énumère tous les arbres possibles pour les phylogénies allant de 2 à 5 espèces (sans tenir compte de l'ordre des espèces). De cette manière, il est possible de construire un tableau qui illustre le nombre d'arbres en fonction du nombre d'espèces et du nombre de noeuds ancestraux (voir tableau 1.3.1).

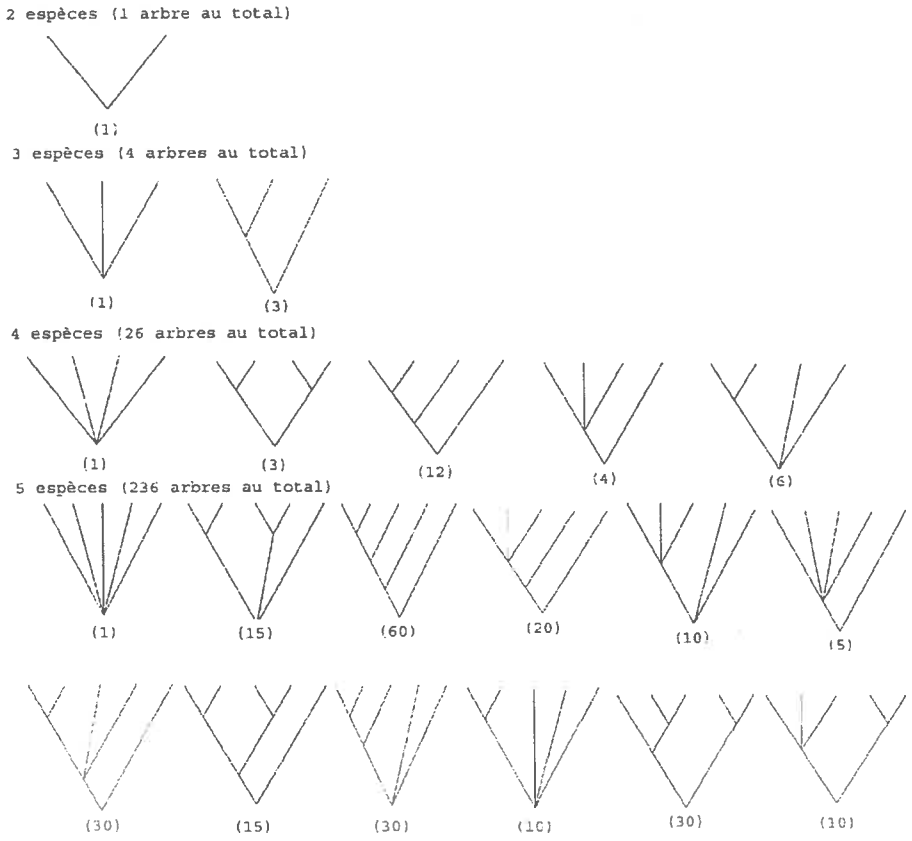


FIGURE 1.3.4. Les arbres polyfurcaux de 2, 3, 4 et 5 espèces (le nombre de permutations des espèces possibles est indiqué entre parenthèses)

À partir d'une démonstration par induction, il est possible de démontrer que le nombre d'arbres suit la formule suivante :

Théorème 1.3.1. *Soit $T_{n,m}$ le nombre d'arbres possibles lorsque nous avons* n *espèces et* m *noeuds ancestraux, alors*

$$T_{n,m} = \begin{cases} (n+m-2)T_{n-1,m-1} + mT_{n-1,m} & si\ m > 1; \\ T_{n-1,m} & si\ m = 1. \end{cases} \tag{1.3.1}$$

DÉMONSTRATION. Le nombre d'arbres augmente avec le nombre d'espèces. Il croît de manière récursive.

Pour la première partie de la preuve par induction, prenons une phylogénie à 2 espèces. Un seul arbre est possible, celui qui rattache les deux espèces à partir d'un noeud ancestral. Ensuite, si nous voulons passer à une phylogénie à trois espèces, il existe deux avenues possibles :

(1) conserver le même nombre de noeuds, soit dans ce cas-ci, un (voir figure 1.3.5) ;

(2) rajouter un noeud, donc ce qui rend le compte à deux noeuds (voir figure 1.3.5).

Dans le premier cas, où nous conservons le même nombre de noeuds, nous avons encore un arbre, puisque nous pouvons connecter la nouvelle espèce qu'à un seul noeud. Dans le deuxième cas où nous voulons rajouter un noeud, nous devons placer notre nouvelle espèce sur l'une des branches de l'arbre précédent. L'arbre précédent compte deux branches. Donc, il est possible de rajouter un noeud sur l'une de ses deux branches. De plus, il est toujours possible de conserver ce même arbre et de lui rajouter un noeud ancestral, ce qui donne un autre arbre possible. Ainsi, nous pouvons créer trois arbres en rajoutant un noeud.

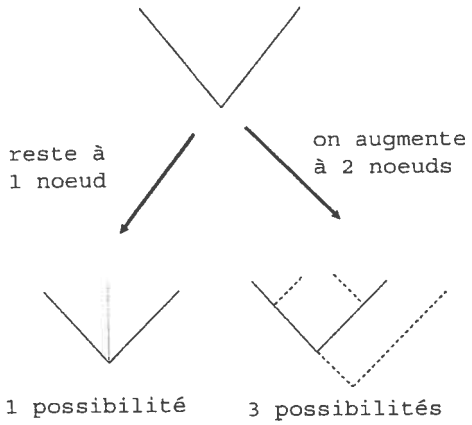


FIGURE 1.3.5. Le passage de 2 espèces à 3 espèces

Pour la deuxième partie de la preuve par induction, prenons une phylogénie à *n-1* espèces et *m-1* noeuds. Ensuite, si nous voulons passer à une phylogénie à n espèces, il existe encore deux avenues possibles :

(1) conserver le même nombre de noeuds, soit dans ce cas-ci, *m-1* noeuds ;

(2) rajouter un noeud, ce qui donne m noeuds.

Dans le premier cas, où nous conservons le même nombre de noeuds, nous pouvons faire le même nombre d'arbres qu'à l'itération précédente fois le nombre de noeuds. Donc, dans ce cas-ci où nous restons à *m-1* noeuds, nous pouvons placer la nouvelle espèce sur l'un des *m-1* noeuds. Donc, ceci donne $(m-1)T_{n,m-1}$. Dans le deuxième cas où nous voulons rajouter un noeud, nous devons placer notre nouvelle espèce sur l'une des branches de l'arbre précédent. De cette manière, le nombre d'arbres est égal au nombre de branches de l'arbre précédent multiplié par le nombre d'arbres existant auquel nous avons rajouté un noeud. De plus, il est toujours possible de conserver le même arbre et rajouter un noeud ancestral à cet arbre, ce qui donne un arbre de plus. Tel qu'expliqué par Felsenstein (2004), le nombre de branches est égal à la somme du nombre d'espèces et du nombre de noeuds moins 1. Toutefois, comme nous voulons rajouter un noeud à l'arbre précédent, qui possède une espèce en moins et un noeud en moins, nous avons donc $(n-1)+(m-1)-1$ branches. Ainsi, en sommant le nombre de branches de l'arbre précédent et l'arbre du noeud ancestral, nous obtenons $(n-1)+(m-1)-1+1 = n+m-2$ de possibilités de plus qu'à l'arbre précédent. Donc, ceci donne $(n+m-2)T_{n-1,m-1}$ arbres possibles.

Dans l'équation (1.3.1), nous voulons connaître le nombre d'arbres pour $T_{n,m}$. Ainsi, $T_{n,m}$ est construit à partir de l'arbre à n-1 espèces et m-1 noeuds, $T_{n-1,m-1}$ (tel que vu dans le deuxième cas) et l'arbre à $T_{n-1,m}$. Si nous n'ajoutons pas de noeud à l'arbre $T_{n-1,m}$, nous avons m noeuds où nous pouvons connecter la nouvelle espèce à l'arbre. Donc, ce qui donne $mT_{n-1,m}$ arbres possibles.

En sommant $(n+m-2)T_{n-1,m-1}$ et $mT_{n-1,m}$, nous obtenons la première partie de l'équation (1.3.1). En ce qui concerne la seconde partie, il faut spécifier que pour $m = 1$, l'équation devient $T_{n-1,m}$, puisque si $m = 1$, alors

TABLEAU 1.3.2. Énumération du nombre d'arbres en fonction du nombre de noeuds et d'espèces

	nombre de noeuds ancestraux							
		1	2	3	4	5	6	nombre total d'arbres
nombre	2	1						1
d'espèces	3	1	3					4
	4	1	10	15				26
	5	1	25	105	105			236
.....	6	1	56	490	1260	945		2752
	7	1	119	1918	9450	17325	10395	39208

$$\begin{aligned} T_{n,m} &= (n+m-2)T_{n-1,m-1} + mT_{n-1,m} \\ &= (n+1-2)T_{n-1,1-1} + 1T_{n-1,1} \\ &= (n-1)T_{n-1,0} + T_{n-1,1} \\ &= (n-1)0 + T_{n-1,1} \\ &= T_{n-1,1} \\ &= T_{n-1,m}. \end{aligned}$$

□

À l'aide du théorème 1.3.1, nous pouvons alors compléter le tableau 1.3.1, ce qui donne le tableau 1.3.2.

L'énumération des arbres représente une partie importante de la recherche. Le nombre élevé d'arbres possibles pour un jeu de données justifie l'utilisation d'une approche bayésienne. Les ordinateurs n'étant pas suffisamment performants, il est impossible, dans un temps raisonnable, de calculer la vraisemblance pour tous les arbres (voir tableau 1.3.3).

TABLEAU 1.3.3. Énumération du nombre d'arbres en fonction du nombre d'espèces

nombre d'espèces	nombre d'arbres
2	1
3	4
4	26
5	236
6	2 752
7	39 208
8	660 032
9	12 818 912
10	282 137 824
11	6 939 897 856
12	$1{,}8867 \times 10^{11}$
13	$5{,}6174 \times 10^{12}$
14	$1{,}8179 \times 10^{14}$
15	$6{,}3537 \times 10^{15}$
16	$2{,}3851 \times 10^{17}$
17	$9{,}5710 \times 10^{18}$
18	$4{,}0884 \times 10^{20}$
19	$1{,}8522 \times 10^{22}$
20	$8{,}8709 \times 10^{23}$
30	$7{,}0717 \times 10^{41}$
40	$1{,}9037 \times 10^{61}$
50	$6{,}85 \times 10^{81}$
100	$3{,}3388 \times 10^{195}$

1.4. LES MÉTHODES DE RECONSTRUCTION

Comment trouver l'arbre phylogénétique d'un groupe d'espèces ? Comment quantifier un arbre par rapport à un autre ? Quel arbre représente mieux la phylogénie de l'espèce ? Pour répondre à cette question, les biologistes se sont basés

sur l'ouvrage de Darwin (1859), *The Origin of Species*, qui dit que toutes les espèces vivantes sont associées les unes aux autres par un ancêtre commun plus ou moins éloigné dans le temps. Donc, il est possible de dire que plus il existe de similarités morphologiques ou génétiques, plus deux espèces sont dites proches. En partant de ce principe, les biologistes ont développé plusieurs méthodes pour pouvoir quantifier la validité d'un arbre :

(1) méthode par parcimonie ;

(2) méthode par maximum de vraisemblance ;

(3) méthode par une approche bayésienne.

1.4.1. Parcimonie

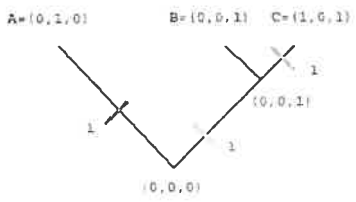
L'analyse par parcimonie se fait par différenciation de caractères homologues. Les caractères peuvent être des nucléotides, ou encore la présence d'organes morphologiques (par exemple des ailes, une corde dorsale, un système digestif, etc.). Donc, si deux espèces n'ont pas le même nucléotide pour un même caractère, nous disons qu'il y a une mutation ou substitution et qu'il y a un «pas» entre les deux espèces. À partir du principe de parcimonie, l'arbre le plus probable sera lui qui compte le moins de «pas», il est donc l'arbre le plus parcimonieux (Darlu et Tassy, 1993).

La figure 1.4.1 et le tableau 1.4.1 donne un exemple d'une analyse d'arbre par parcimonie pour trois espèces (A, B et C). Les «pas» sont représentés par les barres grises. Nous donnons deux arbres et calculons le nombre de «pas» pour chacun par la méthode de parcimonie. À partir du tableau 1.4.1, nous pouvons voir qu'il existe deux différences entre les espèces A et B, trois entre les espèces A et C et une entre les espèces B et C. Pour l'arbre 1, nous devons placer ces différences (mutations), de manière à ce que l'arbre possède le moins de «pas» possible. La manière optimale de faire ceci est de placer un «pas» (mutation) sur la branche qui relie l'espèce A et l'espèce ancestrale, et de placer un autre «pas» sur la branche qui relie l'espèce ancestrale et l'ancêtre commun des espèces B et C. Finalement, en plaçant un autre «pas» sur la branche reliant l'espèce C et l'ancêtre commun des espèces B et C, nous obtenons un arbre à trois «pas». En procédant

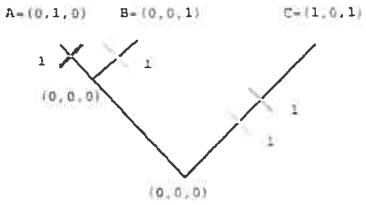
TABLEAU 1.4.1. Tableau des caractères présents ou absents de trois espèces vivantes (0=absence du caractère et 1=présence du caractère)

Espèces \ caractères	1	2	3
A	0	1	0
B	0	0	1
C	1	0	1

de la même manière, il est possible de trouver le nombre de «pas» minimal pour l'arbre 2, qui est de quatre. Alors, l'arbre 1 est le plus parcimonieux. L'arbre phylogénétique de ces trois espèces est l'arbre 1. Toutefois, il faut considérer que nous n'avons pas construit tous les arbres possibles pour ces trois espèces. Donc, il se peut qu'un autre arbre soit plus parcimonieux.



arbre 1



arbre 2

FIGURE 1.4.1. Exemple de parcimonie

Les algorithmes de parcimonie ont été conçus à la fin des années 1960-début des années 1970. Leur efficacité a été régulièrement améliorée. Toutefois, nous

découvrîmes que l'analyse par parcimonie pouvait souvent donner des solutions équivalentes en termes de différences génétiques («pas»), mais qui étaient des non-sens du point vue biologique. En conséquence, les biologistes ont pris recours à d'autres méthodes, telles que le maximum de vraisemblance.

1.4.2. Maximum de vraisemblance

La méthode par maximum de vraisemblance évalue, en terme de probabilités, l'ordre des branchements et la longueur des branches d'un arbre sous un modèle évolutif donné. Depuis les années 1970, les méthodes d'analyse par maximum de vraisemblance sont les plus reconnues, utilisées et développés par les biologistes. Cette méthode sera expliquée plus en détail dans le chapitre 3.

1.5. Conclusion du chapitre

L'analyse phylogénétique des données moléculaires peut se faire par l'entremise de plusieurs méthodes : parcimonie, distance, maximum de vraisemblance, approche bayésienne, etc. Des méthodes non statistiques, telles que l'analyse spectrale sont couramment utilisées. Dans le cadre de ce travail, l'analyse par maximum de vraisemblance et l'analyse par une approche bayésienne seront comparées. Comme Felsenstein (1978) l'a démontré, les méthodes parcimonieuses peuvent faire des reconstructions phylogénétiques incohérentes par rapport à la vraie phylogénie. Ce problème survient surtout lorsque nous sommes en présence d'arbres phylogénétiques possédant de longues branches, la zone de Felsenstein. Donc, pour éviter ce problème, les biologistes travaillent surtout avec l'analyse par maximum de vraisemblance qui fait abstraction des complications provoquées par les longues branches. Depuis cinq ans, le monde de la phylogénétique commence à s'intéresser à la théorie bayésienne qui donne une autre perspective et un nouvel outil à une science qui demeure encore abstraite. Dans le contexte de la phylogénétique, la théorie du maximum de vraisemblance et l'approche bayésienne sont très similaires, les deux emploient un calcul probabiliste pour trouver l'arbre phylogénétique. Toutefois, il demeure une différence entre ces deux méthodes qui réside dans le calcul de la probabilité de l'arbre. Le maximum de vraisemblance

va trouver l'arbre le plus probable d'expliquer les séquences observées $P(D|T)$ (la probabilité des données moléculaires, D, sachant l'arbre phylogénétique, T). Tandis que l'approche bayésienne donnera l'arbre, ou un consensus d'arbres, qui est le plus probable d'être expliqué par les séquences observées $P(T|D)$ (la probabilité des hypothèses phylogénétiques, T, sachant les données moléculaires, D). Cependant, ce problème qui peut sembler anodin, engendre une toute nouvelle méthode d'analyse phylogénétique possédant énormément d'avantages. Nous allons voir et étudier l'analyse par une approche bayésienne dans le chapitre 2.

Chapitre 2

ARBRES PHYLOGÉNÉTIQUES : REPRÉSENTATION ET VRAISEMBLANCE

2.1. Maximum de vraisemblance

L'inférence de l'histoire évolutive des espèces présentée dans ce chapitre repose sur un raisonnement probabiliste. Au début des années 1960, Edwards et Cavalli-Sforza (1963) et Neyman (1971) introduirent en phylogénétique les premières applications des théories de R.A. Fisher (1912) sur le maximum de vraisemblance. Aujourd'hui, une grande partie des études effectuées en phylogénétique se base sur la théorie du maximum de vraisemblance. Elle est ainsi devenue la référence objective en phylogénétique. La méthode du maximum de vraisemblance en phylogénétique consiste à trouver l'arbre le plus probable pour un jeu de données. Mais comment attribuer une probabilité (vraisemblance) à un arbre ?

2.1.1. Vraisemblance d'un arbre

Un arbre est composé de branches et de noeuds, qui sont les espèces vivantes et ancestrales. Toutefois, la probabilité d'un arbre dépend également du nombre de sites (nucléotides).

Hypothèses 2.1.1. *Pour pouvoir calculer la probabilité d'un arbre, nous devons faire deux hypothèses (Swofford, Thorne, Felsenstein et Hillis, 1996b) :*

(1) l'évolution sur chacun des sites (nucléotides) est indépendante ;

(2) l'évolution dans chaque lignée est indépendante.

Sans ces hypothèses, il serait très compliqué, voir impossible, de calculer la probabilité d'un arbre.

Définition 2.1.1. *La probabilité d'un arbre est donnée par l'équation :*

$$L = \Pr(D|T) = \prod_{i=1}^{m} \Pr(D^{(i)}|T), \tag{2.1.1}$$

où D *représente les données,* T *l'arbre pour lequel nous calculons la probabilité et* m *le nombre de sites (nucléotides) de notre séquence d'ADN.*

Posons $L_{(i)} = \Pr(D^{(i)}|T)$. L'équation (2.1.1) peut donc s'écrire sous la forme :

$$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times ... \times L_{(m)} = \prod_{i=1}^{m} L_{(i)}$$

où $L_{(i)}$ est la vraisemblance pour le nucléotide i. Ce produit devient une somme lorsque nous prenons le logarithme naturel de L :

$$\log L = \log L_{(1)} + \log L_{(2)} + \log L_{(3)} + ... + \log L_{(m)} = \sum_{i=1}^{m} \log L_{(i)}.$$

De cette manière, il devient possible d'attribuer une probabilité à tous les arbres. L'arbre avec la plus grande probabilité est le maximum de vraisemblance, donc, l'arbre le plus probable.

2.1.1.1. *Exemple d'un calcul de probabilité pour un arbre*

L'exemple de la figure 2.1.1 nous permet d'illustrer le calcul de la probabilité pour un arbre. Pour le site i, la fonction de probabilité de l'arbre s'écrit :

$$\begin{aligned} L_{(i)} &= \Pr(D^{(i)}|T) \\ &= \sum_x \sum_y \sum_z \sum_w \Pr(x) \Pr(y|x, t_6) \Pr(A|y, t_1) \\ &\quad \times \Pr(C|y, t_2) \Pr(z|x, t_8) \Pr(C|z, t_3) \Pr(w|z, t_7) \\ &\quad \times \Pr(C|w, t_4) \Pr(G|w, t_5), \end{aligned}$$

où x, y, z et w sont des espèces ancestrales inconnues qui peuvent prendre la forme moléculaire du nucléotide de l'adénine (A), la cytosine (C), la guanine (G) ou la thymine (T) sur le site i, tandis que t_k représente la longueur de la branche k de l'arbre.

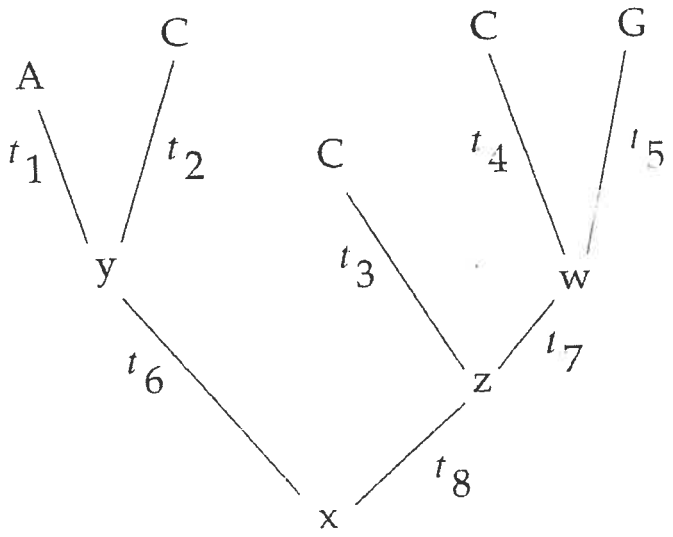


FIGURE 2.1.1. Arbre pour lequel nous calculons la vraisemblance (Felsenstein, 2004)

Donc, à partir de cette formule nous comprenons un peu mieux le calcul de la probabilité d'un arbre. Toutefois, il nous manque encore un élément important pour connaître les valeurs de l'équation. Comment calculer les changements évolutifs ? Ces changements dépendent de plusieurs hypothèses évolutives dictées par les processus de substitution des nucléotides. Il manque un modèle d'évolution des séquences d'ADN permettant d'indiquer si une longueur de branche est appropriée ou si les événements de mutation sont probables.

2.2. UN ARBRE MATRICIEL

Tel que nous l'avons vu dans la section 1.3, un arbre phylogénétique est un graphe permettant d'établir les liens généalogiques entre les espèces. Théoriquement, un arbre est une représentation graphique simple de branches et de noeuds. Toutefois, la transition entre la théorie et la pratique s'avère être une étape cruciale et ardue de la recherche. La question que le chercheur doit se poser est la suivante : comment représenter de manière mathématique un arbre ? Notons qu'il faut se rappeler que nous voulons attribuer une probabilité à un arbre.

Cette section permet de comprendre un peu mieux le processus du calcul informatique de la probabilité conférée à un arbre. Il peut paraître intuitif de

calculer une probabilité, toutefois, lorsqu'il faut mesurer la probabilité pour un graphique, certaines embûches peuvent apparaître. Tout d'abord, il faut définir un arbre sous une forme mathématique. Pour ce faire, nous avons créé pour une représentation matricielle des arbres. Une matrice de 0 et 1 illustre les endroits où les espèces sont liées par des noeuds. Le chiffre 1 symbolise une liaison (un noeud) entre deux espèces voisines et le 0 représente l'absence de liaison entre les deux espèces voisines.

La taille de la matrice dépend du nombre d'espèces dans le jeu de données. Les lignes représentent le temps d'évolution et les colonnes sont les noeuds de liaison entre deux espèces voisines. Par exemple,

$$M = \begin{pmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{pmatrix}$$

est une matrice pour 5 espèces, le nombre de lignes et le nombre de colonnes sont de n-1 espèces. L'élément a représente le noeud au temps 1 entre l'espèce qui prend l'emplacement 1 et l'espèce qui prend l'emplacement 2. Tandis que k défini le noeud au temps 3 entre l'espèce qui prend l'emplacement 3 et l'espèce qui prend l'emplacement 4. Les espèces peuvent permuter, donc, ainsi ils ne sont pas toujours dans l'ordre prédéfini au début de la simulation, c'est pour cela que nous parlons d'emplacements et non d'espèces. Ceci étant dit, une liaison qui se crée entre deux espèces au moment 1, continuera d'exister par la suite. Donc, tous les autres moments qui suivent prendront la valeur 1.

La matrice suivante nous sert d'exemple :

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \tag{2.2.1}$$

Cette matrice est représentée par l'arbre de la figure 2.2.1. Nous pouvons permuter les espèces de l'arbre, donc la matrice (2.2.1) est la matrice mère à 120 arbres.

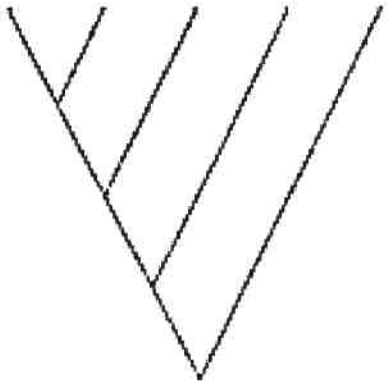


FIGURE 2.2.1. L'arbre représenté par la matrice M (2.2.1)

Cette représentation matricielle nous permet de définir tous les arbres possibles, mais elle n'est pas la clé au calcul probabiliste d'un arbre. Dans la prochaine section, nous abordons le sujet important du catalogage des arbres, puisque le plus gros problème de la phylogénétique est le nombre élevé d'arbres qui existe pour un nombre d'espèces déterminé. Ainsi, pour pouvoir se promener dans l'espace des arbres, il faut répertorier ou avoir un algorithme pouvant construire tous les arbres possibles.

2.3. CATALOGUE DES ARBRES

Comme nous l'avons expliqué dans la section 2.2, le développement d'un algorithme permettant de construire tous les arbres est nécessaire pour poursuivre cette recherche. L'algorithme fonctionne de manière récursive à partir de la représentation matricielle d'un arbre. Nous pouvons construire tous les arbres en partant avec l'arbre à deux espèces (voir figure 2.3.1).

La représentation matricielle du seul arbre pouvant être construit pour une phylogénie à deux espèces est la suivante :

$$M_{21} = \begin{pmatrix} 1 \end{pmatrix}.$$

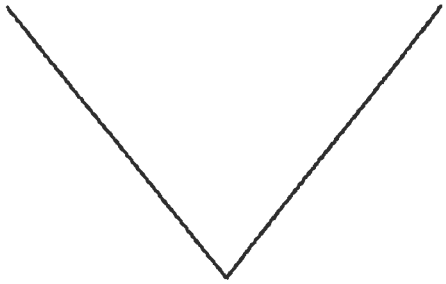


FIGURE 2.3.1. L'arbre représenté par la matrice M_{21}

Pour pouvoir construire les arbres pour une phylogénie de trois espèces, nous devons ajouter une colonne (une nouvelle espèce) à la matrice M_{21}. Donc, si nous rajoutons une espèce à la figure 2.3.1, nous avons deux possibilités :

(1) elle s'attache au même ancêtre que les deux premières espèces (voir la figure 2.3.2 et la matrice (2.3.1)) ;

(2) elle s'embranche à un nouvel ancêtre qui rejoint l'ancêtre des deux premières espèces (voir figure 2.3.3 et la matrice (2.3.2)).

Ainsi, les matrices des deux arbres sont les suivantes :

$$M_{31} = \begin{pmatrix} 1 & 1 \end{pmatrix}, \tag{2.3.1}$$

$$M_{32} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}. \tag{2.3.2}$$

De cette manière, il est possible de faire une itération pour pouvoir construire les arbres pour une phylogénie à trois espèces à partir de l'arbre obtenu pour la phylogénie à deux espèces. Nous devons rajouter une colonne à la matrice M_{21} : cette colonne représente le nouveau noeud. La nouvelle espèce doit se rattacher à

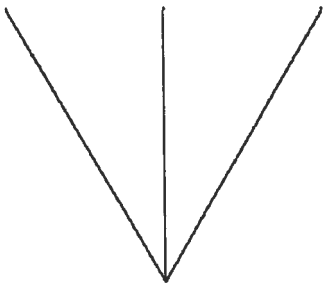


FIGURE 2.3.2. L'arbre représenté par la matrice M_{31}

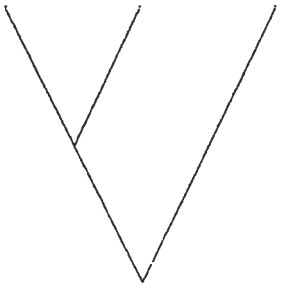


FIGURE 2.3.3. L'arbre représenté par la matrice M_{32}

l'ancêtre des deux premières espèces ou peut s'embrancher à un nouvel ancêtre. Dans le premier cas, l'arbre peut-être représenté par 1 sur la première ligne de la nouvelle colonne, ce qui donne la matrice M_{31} (voir la matrice (2.3.1)). Dans le deuxième cas, la matrice prendra un 0 sur la première ligne de la nouvelle colonne. Ensuite, nous devons rajouter une nouvelle ligne puisque l'arbre n'est pas terminé. En effet, il y a un noeud qui rattache l'espèce 1 et 2, mais l'espèce 3 n'est pas encore rattachée. Ainsi, il faut créer un nouvel ancêtre pour joindre l'arbre. Conséquemment, il est inutile de rajouter un 0 à la nouvelle ligne, puisqu'il donnerait la même information que la première ligne. Nous devons rajouter un 1 à la nouvelle ligne de la nouvelle colonne (voir la matrice (2.3.2)). De cette manière, nous pouvons construire toutes les topologies pour le nombre d'espèces que nous désirons. Il est important à noter que cet algorithme ne fait que créer les topologies existantes, il ne construit pas tous les arbres. Il donne toutes les topologies sans étiquette, c'est-à-dire que pour chacun des noeuds externes, il est possible d'assigner une des espèces du jeu de données. Pour obtenir tous les

arbres, nous devons prendre ces topologies et faire toutes les permutations au niveau des noeuds externes de l'arbre.

Voici un résumé des itérations pour trouver le nombre d'arbres pour la phylogénie à 4 espèces.

La matrice M_{31} donne les deux topologies suivantes :

$$M_{41} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix},$$

$$M_{42} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

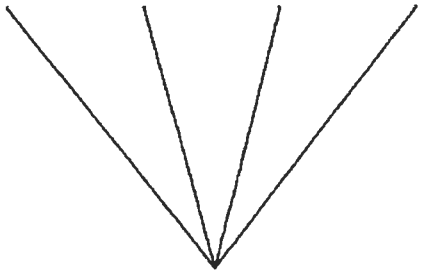


FIGURE 2.3.4. L'arbre représenté par la matrice M_{41}

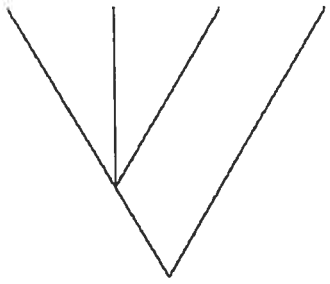


FIGURE 2.3.5. L'arbre représenté par la matrice M_{42}

La matrice M_{32} donne les trois topologies suivantes :

$$M_{43} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

$$M_{44} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix},$$

$$M_{45} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

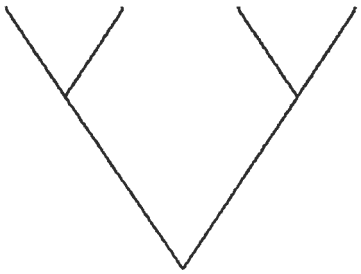


FIGURE 2.3.6. L'arbre représenté par la matrice M_{43}

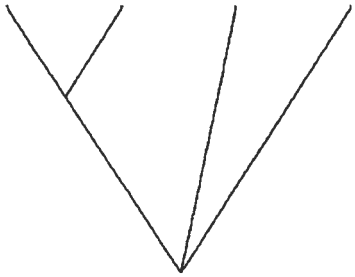


FIGURE 2.3.7. L'arbre représenté par la matrice M_{44}

2.4. CALCUL DE LA PROBABILITÉ POUR UN ARBRE

Notre méthode pour calculer la probabilité pour un arbre est de passer par une multiplication matricielle. Cette méthode permet de systématiser le processus du calcul de la probabilité pour un arbre. D'autres méthodes existent, telle la méthode du *pruning* de Felsenstein (2004).

Tel qu'expliqué dans la sous-section 2.1.1, nous calculons la vraisemblance en sommant le logarithme de la probabilité des sites. À l'intérieur de chaque

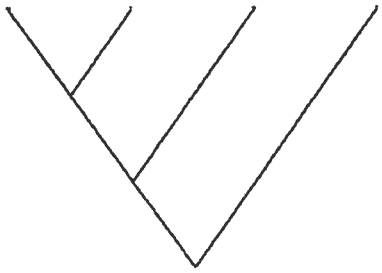


FIGURE 2.3.8. L'arbre représenté par la matrice M_{45}

site, le calcul prend la forme d'un produit des branches, en supposant l'indépendance conditionnelle des branches. Ainsi, il faut reconstruire l'arbre sous forme de branches, puisque chaque branche possède sa probabilité. S'il y a une mutation sur la branche, alors une probabilité correspondante à une mutation sera attribuée à cette branche. Dans le cas de l'absence de mutation, alors la branche prend une autre probabilité. Chaque modèle de substitution a ses propres probabilités, donc ceci complique la tâche. Notre idée est de construire un tableau qui comprendra toutes les possibilités pour chacune des espèces dans un modèle de substitution prédéfini. Pour les fins de cet exemple, nous prenons le modèle de Jukes-Cantor (voir section 3.2.1) afin de construire un tableau de calcul, noté U.

Le calcul d'une probabilité est obtenu avec l'équation suivante :

$$\log L = \log L_{(1)} + \log L_{(2)} + \log L_{(3)} + ... + \log L_{(m)} = \sum_{i=1}^{m} \log L_{(i)}.$$

Nous voulons décomposer la probabilité de l'arbre en un produit matriciel de la forme suivante :

$$\log L = \sum_{k=1}^{m} \log(W(U^{(k)}V)).$$

Soit U un tableau à trois dimensions :

(1) $n = 4^{\text{nombre d'espèces ancestrales}}$ (dimension 1, les lignes du tableau U) ;

(2) s (nombre de substitutions possibles du modèle) $\times$ nombre de branches (dimension 2, les colonnes du tableau U) ;

(3) $m =$ nombre de sites (dimension 3).

Le vecteur V permet d'inclure un modèle biologique (voir section 3.1) au calcul de la probabilité et

$$W = \begin{pmatrix} 1 & 1 & 1 & \dots & \dots & 1 & 1 & 1 \end{pmatrix}^t_{n\times 1}.$$

Le nombre de lignes dans le tableau U est le nombre de nucléotides qui peut être attribué aux ancêtres. Comme il y a quatre nucléotides A, C, G et T, un arbre peut être $4^{\text{nombre d'espèces ancestrales}}$ pour chacun des sites. Le nombre de colonnes dépend du modèle de substitution sélectionné, s est la variable qui donne le nombre de possibilités que peut prendre une branche (mutation, absence de mutation, transition, etc.). Dans le cas simple, où nous sommes en présence d'un modèle de Jukes-Cantor (voir section 3.2.1), $s = 2$, la mutation donne une probabilité, et l'absence de mutation donne une autre probabilité. Toutefois, dans d'autres modèles tel le Kimura à deux paramètres (voir section 3.2.2), la probabilité d'une branche à trois alternatives de sélection, l'absence de mutation, la transition ou la transversion, donc $s = 3$. Finalement, la troisième dimension du tableau U représente le nombre de sites du jeu de données.

Le tableau U permet de calculer la probabilité pour un arbre. L'idée est de compléter le tableau de 0 et 1 de la manière suivante. Prenons l'exemple d'un arbre à 2 espèces ancestrales, 4 branches, 1 site sous le modèle de Jukes-Cantor (voir figure 2.4.1). La matrice (2.4.1) permet d'illustrer le processus de la construction du tableau U.

Les lignes de la matrice $U^{(k)}$, voir équation (2.4.1), représentent les combinaisons possibles que nous pouvons inférer à l'arbre. Ainsi les deux ancêtres de l'exemple (espèce 4 et espèce 5) pourrait prendre les nucléotides (A,A), (A,C), (A,G), (A,T), (C,A), (C,C), (C,G), (C,T), (G,A), (G,C), (G,G), (G,T), (T,A), (T,C), (T,G), ou finalement (T,T), donc ce qui donne $4^{\text{nombre d'espèces ancestrales}}$ possibilités de valeurs inférées aux ancêtres. Ensuite, chaque colonne fonctionne en paire, la première colonne représente une mutation pour la branche t_1 et la deuxième colonne représente une absence de mutation pour la branche t_1, la troisième colonne représente une mutation pour la branche t_2 et la quatrième colonne représente une absence de mutation pour la branche t_2, et ainsi de suite. Dans notre exemple, si nous prenons les nucléotides A et A (voir la figure 2.4.2 et

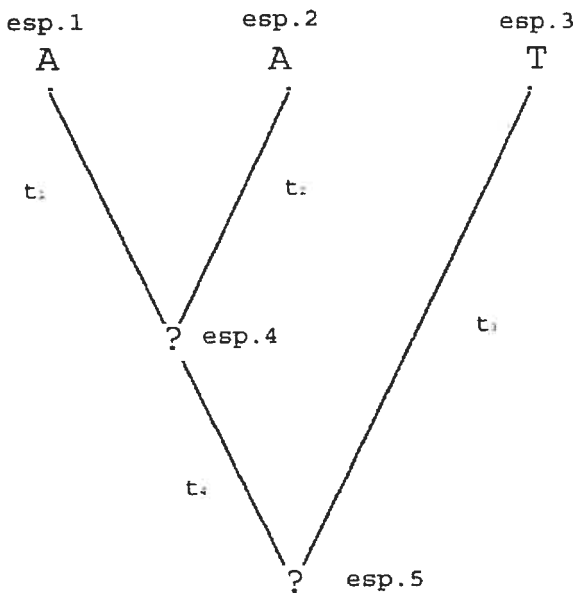


FIGURE 2.4.1. Arbre phylogénétique et les valeurs ancestrales inconnues

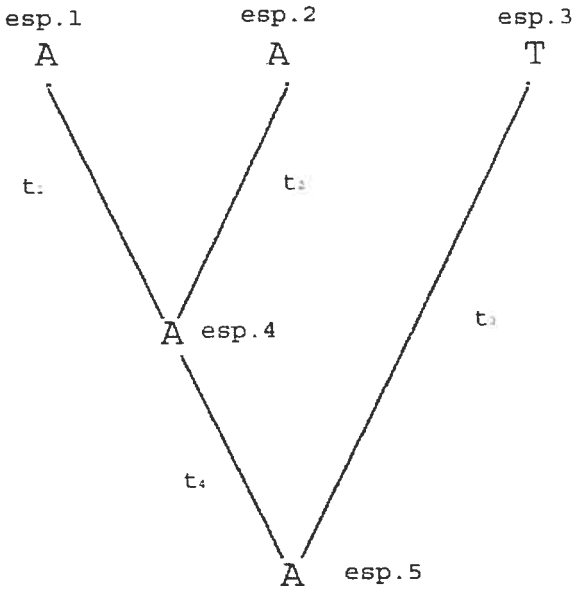


FIGURE 2.4.2. Arbre phylogénétique avec des valeurs inférées (A et A)

la première ligne de la matrice (2.4.1)) pour les espèces 4 et 5, nous avons aucune mutation sur les branches t_1, t_2 et t_4, tandis qu'il y a une mutation pour la branche t_3. La deuxième ligne de la matrice $U^{(k)}$ représente l'arbre où les deux ancêtres, les espèces 4 et 5, prennent les nucléotides C et A, respectivement (voir figure 2.4.3). Donc, pour ce jeu de données nous avons des mutations pour toutes

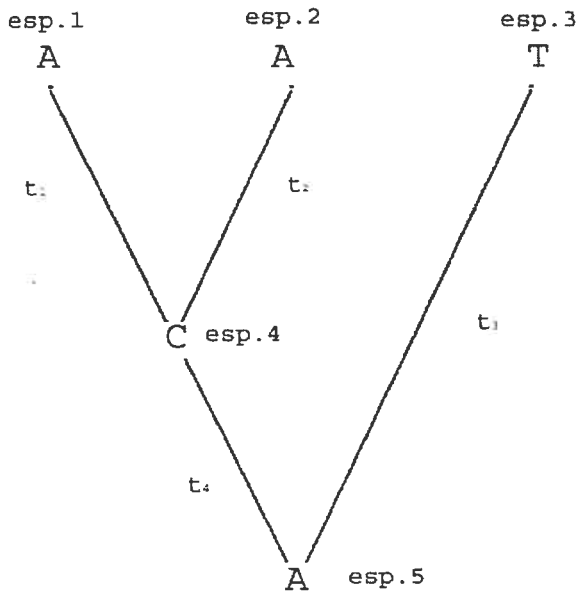


FIGURE 2.4.3. Arbre phylogénétique avec des différentes valeurs inférées (A et C)

les branches, t_1, t_2, t_3 et t_4 de l'arbre. Nous pouvons ainsi compléter la matrice (2.4.1). Nous devons spécifier que les données proviennent du site k, alors nous sommes en présence d'une matrice $U^{(k)}$ du tableau U.

$$U^{(k)} = \left(\begin{array}{cc|cc|cc|cc} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{array}\right). \tag{2.4.1}$$

Ensuite, nous multiplions la matrice $U^{(k)}$ par la vecteur V qui dépend du modèle de substitution utilisé. Ici dans le cas où nous prenons le modèle Jukes-Cantor, le vecteur V prend la forme suivante :

$$V = \begin{pmatrix} \frac{1}{4} - \frac{1}{4}e^{-\mu t_1} \\ \frac{1}{4} + \frac{3}{4}e^{-\mu t_1} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t_2} \\ \frac{1}{4} + \frac{3}{4}e^{-\mu t_2} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t_3} \\ \frac{1}{4} + \frac{3}{4}e^{-\mu t_3} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t_4} \\ \frac{1}{4} + \frac{3}{4}e^{-\mu t_4} \end{pmatrix}.$$

Ainsi, ce vecteur permet d'inclure le modèle de substitution dans la matrice $U^{(k)}$ du tableau U. L'indice de t représente la branche sur laquelle la substitution se produit. Finalement, il ne nous reste qu'à sommer les logarithmes naturels de toutes les sommes des produits $U^{(k)}V$ pour chacun des sites du jeu de données.

2.5. Conclusion du chapitre

Tel que nous l'avons vu dans ce chapitre, l'analyse phylogénétique demande le développement d'un processus mathématique et informatique. À partir des formules pour calculer la vraisemblance des arbres, nous avons construit des algorithmes et des représentations matricielles pour schématiser les arbres. Toutefois, il nous reste maintenant à inclure les modèles biologiques à ce processus. En vérifiant nos équations, nous pouvons nous rendre à l'évidence qu'il nous manque encore un élément important pour connaître les vraisemblances des arbres. Comment pouvons-nous calculer les changements évolutifs (les substitutions, les mutations) ? Ces changements dépendent de plusieurs hypothèses évolutives dictées par les processus de substitution des nucléotides. Dans le prochain chapitre, nous allons discuter des modèles d'évolution pour les séquences d'ADN qui nous permettent de construire des arbres avec des longueurs de branche appropriées et de

vérifier si les événements de mutation sont probables ou non. Notez que la longueur des branches est proportionnelle au temps d'évolution entre deux espèces.

Chapitre 3

MODÈLES DE SUBSTITUTION

3.1. MODÈLES DE SUBSTITUTION : MODÈLE D'ÉVOLUTION DES SÉQUENCES D'ADN

Les modèles de substitution permettent aux méthodes de maximum de vraisemblance et à l'approche bayésienne d'élaborer un modèle statistique sur les arbres. Ces modèles se basent entièrement sur les modèles de Markov, dans lesquels une probabilité de transition entre un état i et un état j sur un site donné ne dépend pas de l'historique du site avant l'obtention de son état i. Nous devons également supposer que la probabilité de substitution ne change pas dans les différentes parties de l'arbre (Felsenstein, 1981 ; Lanave *et al.*, 1984 ; Tavaré, 1986).

Définition 3.1.1. *L'expression mathématique d'un modèle de substitution est un tableau de taux (substitution sur un site par unité de distance évolutive) sur lequel tous les nucléotides peuvent être remplacés par un nucléotide alternatif. Alors, pour une séquence d'ADN, ceci est représenté par une matrice quatre par quatre, notée Q, dans laquelle chaque élément Q_{ij} est un taux de changement d'un nucléotide* i *à un nucléotide* j. *La matrice prend la forme suivante :*

$$Q = \begin{pmatrix} Q_{AA} & Q_{AC} & Q_{AG} & Q_{AT} \\ Q_{CA} & q_{CC} & Q_{CG} & Q_{CT} \\ Q_{GA} & Q_{GC} & Q_{GG} & Q_{GT} \\ Q_{TA} & Q_{TC} & Q_{TG} & Q_{TT} \end{pmatrix}. \tag{3.1.1}$$

La matrice de transition (3.1.1) donne un aperçu du fonctionnement d'une matrice de substitution. Toutefois, les choses se compliquent lorsque nous introduisons un modèle biologique à cette matrice de transition. Plusieurs éléments peuvent influencer la mutation d'un nucléotide. Par exemple, nous avons :

(1) la fréquence relative des nucléotides dans le génome des espèces ;

(2) le taux de substitution moyen d'un nucléotide par un autre peut varier (une mutation impliquant que des purines ou que des pyrimidines).

Donc, voici la matrice de substitution Q qui tient compte de tous ces facteurs :

$$Q = \begin{pmatrix} -\mu\xi_A & \mu\tau_{AC}\pi_C & \mu\tau_{AG}\pi_G & \mu\tau_{AT}\pi_T \\ \mu\tau_{CA}\pi_A & -\mu\xi_C & \mu\tau_{CG}\pi_G & \mu\tau_{CT}\pi_T \\ \mu\tau_{GA}\pi_A & \mu\tau_{GC}\pi_C & -\mu\xi_G & \mu\tau_{GT}\pi_T \\ \mu\tau_{TA}\pi_A & \mu\tau_{TC}\pi_C & \mu\tau_{TG}\pi_G & -\mu\xi_T \end{pmatrix}, \tag{3.1.2}$$

où chaque lignes (et colonnes) correspondent aux nucléotides A, C, G et T respectivement et $\xi_A = \tau_{AC}\pi_C + \tau_{AG}\pi_G + \tau_{AT}\pi_T$, $\xi_C = \tau_{CA}\pi_A + \tau_{CG}\pi_G + \tau_{CT}\pi_T$, $\xi_G = \tau_{GA}\pi_A + \tau_{GC}\pi_C + \tau_{GT}\pi_T$ et $\xi_T = \tau_{TA}\pi_A + \tau_{TC}\pi_C + \tau_{TG}\pi_G$. Par exemple, la possibilité d'une mutation de A vers C est $\mu\tau_{AC}\pi_C$. Le paramètre μ représente le taux de substitution moyen. Ce taux moyen est modifié par les paramètres $\tau_{AC}, \tau_{AG}, \tau_{AT}, \ldots, \tau_{TC}, \tau_{TG}$ qui correspondent à un taux attribué pour chaque transformation possible d'une base à une autre. Les paramètres π_A, π_C, π_G et π_T sont les paramètres de fréquence relative pour chacune des bases (nucléotides) A, C, G et T, respectivement.

Définition 3.1.2. *La matrice du taux de substitution Q donne le taux de changement entre les paires de bases (nucléotides) pour un instant de temps, noté dt. Toutefois, pour calculer la vraisemblance d'un arbre, nous avons besoin de calculer la probabilité de substitution d'un état* i *à un état* j *sur une branche de longueur t (la longueur de branche est le produit du temps écoulé et d'une constante non déterminée). La matrice de la probabilité de substitution se calcule à partir de la*

fonction suivante :

$$\Pr(t) \propto e^{Qt}, \tag{3.1.3}$$

où Q est la matrice de substitution et t la longueur de la branche (temps évolutif).

L'exponentielle est évaluée en décomposant la matrice du taux de substitution en valeurs propres.

3.1.1. Valeurs propres

Définition 3.1.3. *Soit P, une matrice carrée. Nous disons que le nombre réel λ est une valeur propre de P et que le vecteur colonne X est un vecteur propre de P associé à λ si*

$$X \neq 0,$$

$$PX = \lambda X.$$

3.1.2. Tous les modèles de substitution

Tous les modèles de substitution se basent sur la matrice (3.1.2). Toutefois, il est généralement supposé que le taux global de substitution d'une base (nucléotide) i à une base (nucléotide) j pour une longueur de temps évolutif donnée est le même que le taux de substitution pour un changement d'une base j à une base i. Ces types de modèle sont à temps-réversible. Par conséquent, les paramètres $\tau_{AC}, \tau_{AG}, \tau_{AT}, ..., \tau_{TC}, \tau_{TG}$ de la matrice (3.1.2) peuvent maintenant être simplifiés : $\tau_{CA} = \tau_{AC}$, $\tau_{GA} = \tau_{AG}$, $\tau_{TA} = \tau_{AT}$, $\tau_{GC} = \tau_{CG}$, $\tau_{TC} = \tau_{CT}$ et $\tau_{TG} = \tau_{GT}$. En faisant les changements suivants à la matrice (3.1.2), nous obtenons le modèle général à temps-réversible (GTR, voir la sous-section 3.1.3). Nous pouvons également supposer les fréquences des bases comme étant égales ce qui donne d'autres types de modèles (c'est-à-dire K2P, voir la sous-section 3.2.2). Il est possible d'imposer des restrictions sur les transversions et les transitions, ce qui donne les modèles F81 (voir la sous-section 3.2.3), F84, HKY85, etc. Lorsque nous supposons que les fréquences des bases sont égales et que toutes les substitutions surviennent au même taux ($\tau_{AC} = \tau_{AG} = \tau_{AT} = \tau_{CG} = \tau_{CT} = \tau_{GT} = 1$), nous obtenons le modèle Jukes-Cantor (JC, voir la sous-section 3.2.1). La figure 3.1.1

présente la hiérarchie des modèles, plus le modèle est haut dans le diagramme, plus il est complexe et par conséquent son temps de simulation augmente.

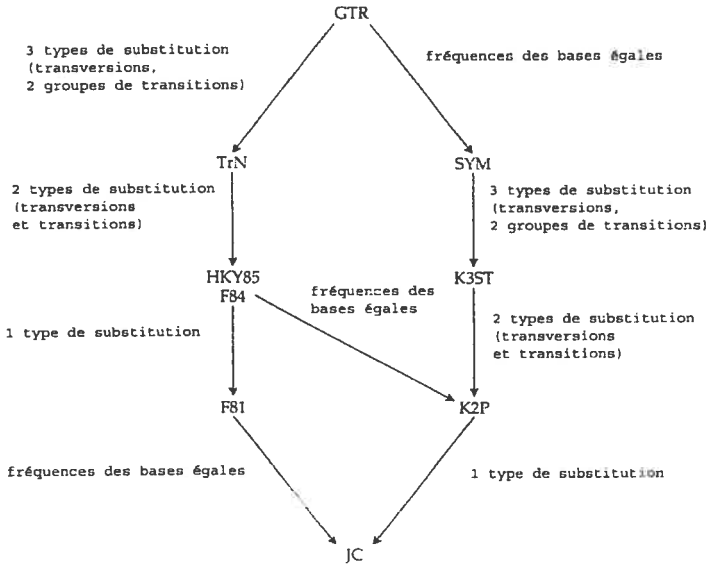


FIGURE 3.1.1. Tous les modèles de substitution (Swofford, Thorne, Felsenstein et Hillis, 1996b)

3.1.3. Modèle général à temps réversible (GTR)

Le modèle général à temps réversible, noté GTR (Lanave *et al.*, 1984 ; Tavaré, 1986 ; Rodriguez *et al.*, 1990), est le modèle le plus complexe possédant un grand nombre de paramètres à estimer (c'est-à-dire 10 + le nombre de branches). Voici sa matrice de transition :

$$Q = \begin{pmatrix} -\mu\xi_A & \mu\tau_{AC}\pi_C & \mu\tau_{AG}\pi_G & \mu\tau_{AT}\pi_T \\ \mu\tau_{AC}\pi_A & -\mu\xi_C & \mu\tau_{CG}\pi_G & \mu\tau_{CT}\pi_T \\ \mu\tau_{AG}\pi_A & \mu\tau_{CG}\pi_C & -\mu\xi_G & \mu\tau_{GT}\pi_T \\ \mu\tau_{AT}\pi_A & \mu\tau_{CT}\pi_C & \mu\tau_{GT}\pi_G & -\mu\xi_T \end{pmatrix},$$

où le paramètre μ représente le taux moyen de substitution moyen et $\xi_A = \tau_{AC}\pi_C+\tau_{AG}\pi_G+\tau_{AT}\pi_T$, $\xi_C = \tau_{AC}\pi_A+\tau_{CG}\pi_G+\tau_{CT}\pi_T$, $\xi_G = \tau_{AG}\pi_A+\tau_{CG}\pi_C+\tau_{GT}\pi_T$ et $\xi_T = \tau_{AT}\pi_A + \tau_{CT}\pi_C + \tau_{GT}\pi_G$. Ce taux moyen est modifié par les paramètres τ_{AC}, τ_{AG}, τ_{AT}, τ_{CG}, τ_{CT}, τ_{GT} qui correspondent à chaque transformation possible

d'une base à une autre. Les paramètres π_A, π_C, π_G et π_T sont les paramètres de fréquence relative pour chacune des bases A, C, G et T, respectivement.

Ce modèle souvent très complexe prend énormément de temps à simuler, ce qui est dû au nombre élevé de paramètres à estimer. Pour cette raison, ce modèle est très peu utilisé. Le nombre d'arbres que nous avons à simuler étant très grand, nous devons tenter de diminuer le temps de simulation le plus possible.

3.2. Modèles les plus utilisés

Incidemment, suite aux difficultés que peut amener le modèle général à temps réversible, la majorité des chercheurs en biologie utilise d'autres modèles tels que le Jukes-Cantor (JC, voir la sous-section 3.2.1), le Kimura à deux paramètres (K2P, voir la sous-section 3.2.2) et le Felsenstein introduit en 1981 (F81, voir la sous-section 3.2.3).

3.2.1. Modèle de Jukes-Cantor (JC)

Tel que nous l'avons expliqué au début du chapitre, lorsque nous supposons les fréquences des bases égales ($\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$) et que tous les taux de substitution sont égaux ($\tau_{AC} = \tau_{AG} = \tau_{AT} = \tau_{CG} = \tau_{CT} = \tau_{GT} = 1$), nous obtenons le modèle Jukes-Cantor, noté JC (Jukes et Cantor, 1969). Ceci donne la matrice Q suivante :

$$Q = \begin{pmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{pmatrix}, \tag{3.2.1}$$

où le paramètre μ représente le taux moyen de substitution.

Théorème 3.2.1. *Si nous posons* $\Pr(t) \propto e^{Qt}$ *où* t *est la longueur des branches et* Q *la matrice de substitution selon le modèle de Jukes-Cantor, alors l'obtention des valeurs propres de l'équation (3.1.2) nous donnera :*

(1) une probabilité de mutation $(= \frac{1}{4} - \frac{1}{4}e^{-\mu t})$ *;*

(2) une probabilité d'absence de mutation $(= \frac{1}{4} + \frac{3}{4}e^{-\mu t})$.

DÉMONSTRATION. Si

$$\Pr(t) \propto e^{Qt},$$

où Q est la matrice (3.2.1) et t la longueur des branches,

alors,

$$\Pr(t) = \begin{pmatrix} e^{-\frac{3}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} \\ e^{\frac{1}{4}\mu t}\mu & e^{-\frac{3}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} \\ e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{-\frac{3}{4}\mu t} & e^{\frac{1}{4}\mu t} \\ e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{-\frac{3}{4}\mu t} \end{pmatrix}$$

et $X(\lambda)$ est solution de

$$\det \begin{pmatrix} e^{-\frac{3}{4}\mu t} - \lambda & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} \\ e^{\frac{1}{4}\mu t}\mu & e^{-\frac{3}{4}\mu t} - \lambda & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} \\ e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{-\frac{3}{4}\mu t} - \lambda & e^{\frac{1}{4}\mu t} \\ e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{\frac{1}{4}\mu t} & e^{-\frac{3}{4}\mu t} - \lambda \end{pmatrix} = 0.$$

À l'aide du logiciel Mathematica 4.0, nous trouvons les valeurs propres suivantes :

$$\lambda_1 = \lambda_2 = \lambda_3 = e^{-\frac{\mu t}{4}}(e^{\mu t} - 1),$$
$$\lambda_4 = e^{-\frac{\mu t}{4}}(e^{\mu t} + 3).$$

En normalisant les valeurs propres, nous obtenons ces valeurs

$$\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{4} - \frac{1}{4}e^{-\mu t},$$
$$\lambda_4 = \frac{1}{4} + \frac{3}{4}e^{-\mu t}.$$

□

De cette manière, nous avons les probabilités pour les substitutions d'une base (i) à une autre (j) sous un modèle de Jukes-Cantor (JC), notées $\Pr_{ij}(t)$ qui sont :

$$Pr_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & i = j; \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & i \neq j; \end{cases} \tag{3.2.2}$$

où le paramètre μ représente le taux moyen de substitution et t la longueur de la branche (temps évolutif).

La matrice des probabilités de substitution prend la forme suivante :

$$\Pr(t) = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} + \frac{3}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} + \frac{3}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} - \frac{1}{4}e^{-\mu t} & \frac{1}{4} + \frac{3}{4}e^{-\mu t} \end{pmatrix}.$$

3.2.2. Modèle de Kimura à deux paramètres (K2P)

Tel que nous l'avons expliqué au début de cette section, il existe un type de modèle qui considère que les fréquences des bases (nucléotides) sont égales. Donc, ce qui veut dire que le génome de toutes les espèces est composé de 25% d'adénine, de 25% de cytosine, de 25% de guanine et de 25% de thymine. Nous savons qu'en réalité les fréquences des bases sont semblables, mais non équivalentes. Toutefois, cette hypothèse permet de réduire le nombre d'estimation de paramètres. Donc, il ne reste que trois probabilités à considérer : la probabilité de substitution en transitions, la probabilité de substitution en transversions et finalement la probabilité d'aucune substitution. Ce modèle, dit Kimura à deux paramètres, noté K2P (Kimura, 1981), ne comporte que deux paramètres libres à évaluer (sans compter les longueurs des branches). La matrice Q du modèle K2P suppose donc que les transversions et les transitions surviennent à différents taux. Ainsi, nous pouvons simplifier la matrice (3.1.3) en fixant $\tau_{AC} = \tau_{AT} = \tau_{CG} = \tau_{GT} = 1$ et $\tau_{AG} = \tau_{CT} = \kappa$ pour obtenir la matrice Q suivante :

$$Q = \begin{pmatrix} -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa \\ \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{1}{4}\mu(\kappa+2) \end{pmatrix},$$

où le paramètre μ représente le taux moyen de substitution et κ est un paramètre qui donne un poids au rapport *transition* : *transversion*.

Un $\kappa = 1$ indique qu'il n'existe pas de différence entre le nombre de transitions et le nombre de transversions, donc nous revenons au modèle de Jukes-Cantor. Toutefois, un $\kappa = 2$ implique que notre modèle suppose qu'il existe deux fois plus de transitions que de transversions (voir figure 1.2.3). En réalité, nous savons qu'il y a deux fois plus de transversions que de transitions, donc, ce paramètre permet d'ajuster ce phénomène.

En utilisant la même méthode des valeurs propres que pour le modèle Jukes-Cantor (JC), voir le théorème 3.2.1, il est possible d'obtenir les probabilités de substitution pour le modèle de K2P. Les probabilités sont les suivantes :

$$Pr_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}e^{-\mu t} + \frac{1}{2}e^{-\mu t(\frac{\kappa+1}{2})} & i = j; \\ \frac{1}{4} + \frac{1}{4}e^{-\mu t} - \frac{1}{2}e^{-\mu t(\frac{\kappa+1}{2})} & i \neq j, \text{ transition}; \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & i \neq j, \text{ transversion}; \end{cases} \tag{3.2.3}$$

où le paramètre μ représente le taux moyen de substitution, κ est un paramètre donnant un poids au rapport *transition* : *transversion* et t la longueur de la branche (temps évolutif).

La matrice des probabilités de substitution prend la forme suivante :

$$\mathrm{Pr}(t) = \frac{1}{4}\begin{pmatrix} 1+e^{-\mu t}+2e^{-\mu t(\frac{\kappa+1}{2})} & 1-1e^{-\mu t} & 1+1e^{-\mu t}-2e^{-\mu t(\frac{\kappa+1}{2})} & 1-1e^{-\mu t} \\ 1-e^{-\mu t} & 1+e^{-\mu t}+2e^{-\mu t(\frac{\kappa+1}{2})} & 1-e^{-\mu t} & 1+e^{-\mu t}-2e^{-\mu t(\frac{\kappa+1}{2})} \\ 1+e^{-\mu t}-2e^{-\mu t(\frac{\kappa+1}{2})} & 1-e^{-\mu t} & 1+e^{-\mu t}+2e^{-\mu t(\frac{\kappa+1}{2})} & 1-e^{-\mu t} \\ 1-e^{-\mu t} & 1+e^{-\mu t}-2e^{-\mu t(\frac{\kappa+1}{2})} & 1-e^{-\mu t} & 1+e^{-\mu t}+2e^{-\mu t(\frac{\kappa+1}{2})} \end{pmatrix}.$$

3.2.3. Modèle de Felsenstein 1981 (F81)

Le modèle de Felsenstein 1981, noté F81 (Felsenstein, 1981), fut développé sous la contrainte où tous les taux de substitution sont égaux, donc $\tau_{AC} = \tau_{AG} = \tau_{AT} = \tau_{CG} = \tau_{CT} = \tau_{GT} = 1$. Toutefois dans ce modèle, ce sont les fréquences relatives des nucléotides qui sont les paramètres libres. Donc, nous obtenons la matrice de substitution Q suivante :

$$Q = \begin{pmatrix} -\mu(\pi_C + \pi_G + \pi_T) & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \pi_T) & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & -\mu(\pi_A + \pi_C + \pi_T) & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & -\mu(\pi_A + \pi_C + \pi_G) \end{pmatrix},$$

où le paramètre μ représente le taux moyen de substitution et π_A, π_C, π_G et π_T représentent respectivement les fréquences relatives des nucléotides A, C, G et T.

En trouvant les valeurs propres à l'équation $\Pr(t) = e^{Qt}$, nous obtenons les probabilités suivantes :

$$Pr_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j)e^{-\mu t} & i = j; \\ \pi_j(1 - e^{-\mu t}) & i \neq j. \end{cases} \tag{3.2.4}$$

La matrice des probabilités de substitution prend la forme suivante :

$$\Pr(t) = \begin{pmatrix} \pi_A + (1 - \pi_A)e^{-\mu t} & \pi_C(1 - e^{-\mu t}) & \pi_G(1 - e^{-\mu t}) & \pi_T(1 - e^{-\mu t}) \\ \pi_A(1 - e^{-\mu t}) & \pi_C + (1 - \pi_C)e^{-\mu t} & \pi_G(1 - e^{-\mu t}) & \pi_T(1 - e^{-\mu t}) \\ \pi_A(1 - e^{-\mu t}) & \pi_C(1 - e^{-\mu t}) & \pi_G + (1 - \pi_G)e^{-\mu t} & \pi_T(1 - e^{-\mu t}) \\ \pi_A(1 - e^{-\mu t}) & \pi_C(1 - e^{-\mu t}) & \pi_G(1 - e^{-\mu t}) & \pi_T + (1 - \pi_T)e^{-\mu t} \end{pmatrix}.$$

3.3. Optimisation de la longueur des branches

Une fois notre modèle de substitution choisi, nous devons calculer la probabilité d'un arbre :

$$L = \Pr(D|T) = \prod_{i=1}^{m} \Pr(D^{(i)}|T) = \prod_{i=1}^{m} L_{(i)},$$

où D représente les données, T l'arbre pour lequel nous calculons la probabilité et m le nombre de sites (nucléotides) de notre séquence d'ADN.

Toutefois, il est impossible de calculer la probabilité d'un arbre seulement avec une topologie. Nous devons décomposer l'arbre en branches et calculer la probabilité de chacune de ces branches. Et comme chaque longueur d'une branche peut en influencer une autre, il devient difficile de calculer une probabilité pour une certaine topologie d'arbre. En d'autres mots, deux arbres avec la même topologie peuvent avoir deux probabilités complètement distinctes (voir figure 3.3.1 et section 2.4). Donc, il existe un espace de probabilité (une densité) pour chaque arbre et chaque jeu de données. Nous devons trouver les longueurs de branches optimales pour notre topologie et jeu de données. En résumé, le chercheur doit, tout d'abord, obtenir la meilleure topologie, et ensuite, trouver les longueurs de branches optimales pour cette topologie.

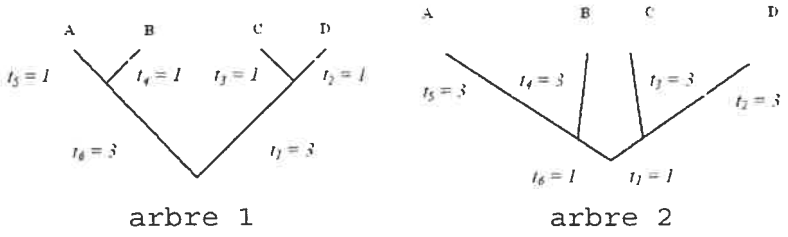


FIGURE 3.3.1. Deux arbres avec la même topologie, mais avec des longueurs de branches différentes

Ceci étant dit, il est impossible de faire suivre une distribution statistique sur les topologies d'arbres et il n'existe pas de solution analytique pour résoudre le problème de l'optimisation de la longueur des branches. Ainsi, nous ne sommes jamais certain d'obtenir l'arbre le plus probable sans les avoir tous calculés. Tel que vu dans la section 1.3.5, le calcul de la probabilité pour chaque arbre prendrait

un temps déraisonnable. Ce problème sera réglé en partie par l'approche bayésienne. Toutefois, le problème de l'absence de solution analytique pour résoudre l'optimisation de la longueur des branches demeure.

Le calcul de la probabilité pour un arbre se fait branche par branche et ensuite site par site. Ainsi, l'optimisation peut se compliquer, puisque si nous optimisons site par site, chaque site aura des longueurs de branche optimales (voir figure 3.3.2). Par conséquent, nous risquons d'obtenir des longueurs de branche différentes d'un site à l'autre. Alors, nous devons conjuguer la maximisation des branches pour tous les sites. Ce problème plus simple et plus communément rencontré dans le domaine de l'optimisation peut être solutionné à l'aide de l'algorithme EM.

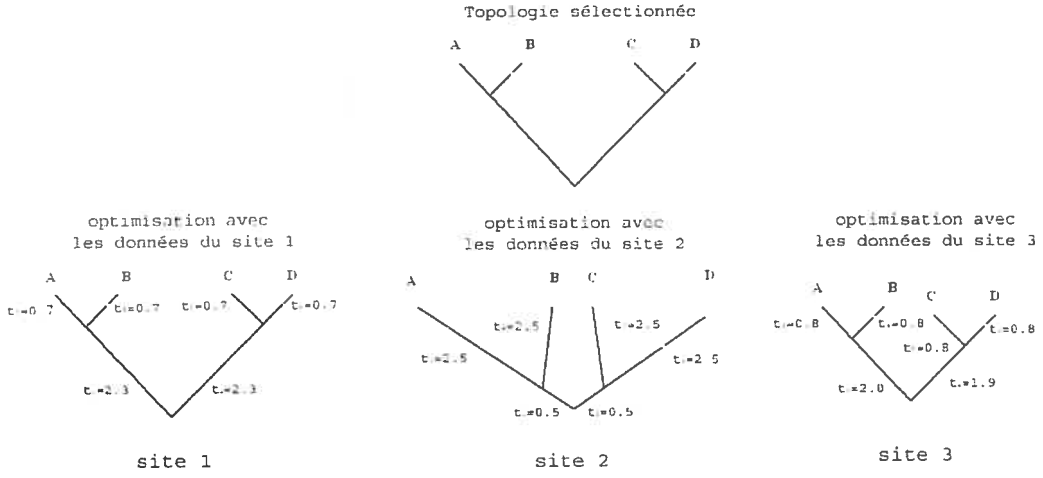


FIGURE 3.3.2. Optimisation site par site

3.3.1. Algorithme EM

Pour estimer les paramètres de la longueur de branches t, nous utilisons la méthode du maximum de vraisemblance. Dans notre cas, il s'agit de maximiser la fonction de vraisemblance $L_{(i)}$ pour chacun des sites. Le système d'équations qui découle de ce problème de maximisation ne peut pas être résolu explicitement,

il faut passer par une méthode itérative telle que l'algorithme EM (voir Dempster, Laird et Rubin, 1977).

Supposons que nous voulons trouver $t = \hat{t}$ pour maximiser la vraisemblance $L = \prod_{i=1}^{m} L_{(i)} = \prod_{i=1}^{m} \Pr(D^{(i)}|T) = \Pr(D|t)$ où D est un jeu de données des sites du gène, T est l'arbre dont nous voulons écrire la probabilité en fonction de ces branches, notée t. À partir d'une approximation initiale $t^{(0)}$, l'algorithme EM génère une suite d'estimés $t^{(m)}$. Chaque itération se compose de deux étapes :

Étape E : Évaluation de $\mathcal{Q}(t, t^{(m)}) = \mathbb{E}[\Pr(D^{(i)}|t)|t^{(m)}]$.

Étape M : Trouver $t = t^{(m+1)}$ qui maximise $\mathcal{Q}(t, t^{(m)})$.

À partir de cet algorithme, il est maintenant possible de calculer la vraisemblance pour un arbre phylogénétique et trouver l'arbre le plus probable selon la méthode du maximum de vraisemblance.

3.4. Conclusion du chapitre

Ce chapitre nous a permis de comprendre le processus des modèles de substitution génétique. En décrivant les modèles Jukes-Cantor, Felsenstein 1981 et le Kimura à deux paramètres, nous avons fait un survol des modèles de substitution. Ces modèles permettent de construire tous les autres modèles (voir figure 3.1.1). Toutefois, il n'est guère intéressant de travailler sur les autres modèles, puisqu'ils sont moins utilisés et possèdent un temps de simulation beaucoup plus élevé. Dans les trois premiers chapitres, nous avons expliqué les bases de la phylogénétique probabiliste. Maintenant, nous devons développer des méthodes statistiques pour pouvoir trouver les arbres les plus probables. Comment allons-nous trouver l'arbre le plus probable parmi une nombre démesuré d'arbres (voir section 1.3.5) ? Nous avons expliqué brièvement l'optimisation de la longueur des branches. Toutefois, dans le prochain chapitre nous allons aborder d'autres méthodes statistiques en phylogénétique. De plus, nous allons introduire l'approche bayésienne dans le domaine de la phylogénétique et les méthodes employées pour obtenir des arbres optimaux selon cette approche. De cette manière, il deviendra possible de comparer la méthode du maximum de vraisemblance et l'approche bayésienne, ce

qui permettra de mieux comprendre les avantages et les désavantages des deux méthodes statistiques dans le domaine de la phylogénétique.

Chapitre 4

UNE APPROCHE BAYÉSIENNE DE LA PHYLOGÉNÉTIQUE

L'approche bayésienne est mathématiquement très similaire à celle du maximum de vraisemblance. Les équations sont très similaires et les modèles de substitution sont les mêmes. Les deux ne diffèrent que dans l'utilisation d'une distribution *a priori* sur les paramètres de l'arbre. L'approche bayésienne utilise la distribution *a priori* dans l'estimation de la topologie des arbres et la longueur des branches. Donc, elle permet aux biologistes de placer plus d'importance sur certains types d'arbres qui leurs semblent plus probables. Ceci permet de sauver énormément de temps de simulation, puisqu'en général les arbres les plus probables sont ceux qui intuitivement le sont. Toutefois, certains biologistes s'opposent à cette idée d'attribuer des poids à certains arbres. Ils critiquent la subjectivité de cette méthodologie. En conséquence, nous pouvons prendre une approche plus neutre, objective, au problème de l'approche bayésienne. Il est possible de donner le même poids à tous les arbres phylogénétiques, tel que pour l'approche du maximum de vraisemblance, et ceci règle le problème de l'objectivité biologique. De plus, dans cette section nous allons voir qu'il y a plusieurs avantages à utiliser l'approche bayésienne.

4.1. Théorème de Bayes

Définition 4.1.1. *Selon le théorème de Bayes, il devient possible de calculer la probabilité conditionnelle* $\Pr(T|D)$ *par l'équation suivante :*

$$si\ \Pr(D) \neq 0, alors\ \Pr(T|D) = \frac{\Pr(T)\Pr(D|T)}{\sum_i \Pr(T_i)\Pr(D|T_i)}, \tag{4.1.1}$$

où T représente un arbre phylogénétique et D les données génétiques.

Pour pouvoir démontrer cette équation nous devons passer par les propriétés des probabilités conditionnelles.

Définition 4.1.2. *En théorie des probabilités, la probabilité conditionnelle d'un événement A sachant qu'un autre B de probabilité non nulle, s'est réalisé est le nombre noté $\Pr(A|B)$ défini par :*

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{4.1.2}$$

Corollaire 4.1.1. *À partir de l'équation (4.1.2), nous obtenons :*

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B),$$

ou encore en inversant A et B dans la probabilité conditionnelle, il est possible d'obtenir une autre équation pour la probabilité de l'intersection.

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)}.$$

Alors,

$$\Pr(B \cap A) = \Pr(B|A)\Pr(A).$$

Définition 4.1.3. *La probabilité d'un événement A peut être calculée par la somme des produits des probabilités conditionnelles de l'événement A sur tous les événements B_i multipliées par la probabilité que l'événement B_i se produise (étant donné que $\Pr(B_i) > 0$, $\sum_i \Pr(B_i) = 1$ et $B_i \cap B_j = \emptyset\ \forall i \neq j$).*

Ainsi,

$$\Pr(A) = \sum_i \Pr(B_i)\Pr(A|B_i).$$

Suite à ces définitions, il est maintenant possible d'intégrer le théorème de Bayes dans le calcul des probabilités des arbres phylogénétiques.

Suite aux définitions et corollaire vus précédemment, nous obtenons le résultat suivant : $\Pr(T|D) = \frac{\Pr(T)\Pr(D|T)}{\sum_i \Pr(T_i)\Pr(D|T_i)}$, où T_i est l'arbre i et D le jeu de données.

4.2. Estimation numérique

Les estimations faites à partir d'une approche bayésienne peuvent devenir compliquées car les calculs en inférence bayésienne sont souvent très long. L'expression pour la distribution *a posteriori* possède un dénominateur

$$\sum_i \Pr(T_i) \Pr(D|T_i)$$

qui peut devenir difficile à calculer, puisque nous avons besoin d'évaluer $\Pr(D|T_i)$ pour tous les arbres possibles. Dans cette sous-section nous allons expliquer les méthodes, algorithmes et propriétés de notre approche bayésienne pour obtenir les arbres phylogénétiques.

4.2.1. MCMC

Lorsque les personnes parlent de méthodes bayésiennes, ils parlent plus souvent de la méthodologie d'estimation MCMC, « Markov chain Monte Carlo » (voir Metropolis *et al.*, 1953), car comme nous l'avons expliqué au début de la sous-section, les calculs en inférence bayésienne peuvent devenir compliqués. Le problème provient du dénominateur de la distribution *a posteriori*, $\sum_i \Pr(T_i) \Pr(D|T_i)$, qui nous demande d'évaluer $\Pr(D|T_i)$ pour tous les arbres possibles. Heureusement, l'échantillonnage sur la densité *a posteriori* peut être obtenu à partir d'une chaîne de Markov. Ainsi, en utilisant des méthodes de Monte-Carlo markoviennes, dites méthodes MCMC, il est maintenant possible de calculer des probabilités bayésiennes dans plusieurs domaines, dont la phylogénétique. Ainsi, la stratégie MCMC cherche un consensus parmi un grand nombre de solutions à un problème. Comme nous le savons, l'un des principaux problèmes en reconstruction phylogénétique est le nombre très grand de solutions possibles pour un jeu de données moléculaires. Le regroupement de dix taxons peut donner plus d'un million d'arbres différents. La méthode MCMC échantillonne parmi ces solutions pour donner celle qui possède la plus grande probabilité *a posteriori* (MAP). De plus, en échantillonnant sur des arbres indépendants, la méthode MCMC donne un intervalle de crédibilité aux phylogénies et un arbre de consensus (arbre moyen), voir section 4.2.4.

4.2.2. L'algorithme Metropolis-Hastings

L'algorithme Metropolis-Hastings (Metropolis *et al.*, 1953; Hastings, 1970) est une méthode permettant d'effectuer l'algorithme du MCMC. L'idée générale derrière l'algorithme MCMC est de se promener dans l'espace des arbres de manière à tomber sur une distribution stationnaire des arbres qui est la distribution *a posteriori*. L'espace des arbres peut être imagé sous la forme d'un grand graphe regroupant tous les arbres connectés les uns aux autres. En supposant que nous connaissons la distribution des arbres *f(t)*, l'algorithme Metropolis-Hastings fonctionne de la manière suivante :

(1) nous posons $i = 0$ et ensuite nous sélectionnons un arbre T_i de départ avec des longueurs de branches aléatoire.

(2) Nous proposons un nouvel arbre ξ en respectant les conditions suivantes :
a) $\Pr(T_i \longrightarrow \xi) = \Pr(\xi \longrightarrow T_i)$;
b) tous les arbres possibles dans la distribution peuvent être atteints si la chaîne de Markov est parcourue assez longtemps ;
c) la chaîne ne doit pas être périodique.

(3) Le nouvel arbre proposé ξ est accepté à la place de T_i avec probabilité $R = \min[1, \frac{\Pr(\xi)}{\Pr(T_i)} \times \frac{\Pr(D|\xi)}{\Pr(D|T_i)}]$.

(4) Nous générons un nombre aléatoire (u) à partir d'une distribution uniforme (0,1) :
si $u \leq R$, alors $T_{i+1} = \xi$;
si $u > R$, alors $T_{i+1} = T_i$.

(5) Nous répétons les étapes 2 à 4 plusieurs fois. Le nombre de fois qu'un arbre T_i est visité par la chaîne MCMC sur le nombre total d'itérations représente une approximation non biaisée de $\Pr(T_i|D)$.

L'algorithme Metropolis-Hastings ne se termine jamais. Une chaîne de Markov peut continuer indéfiniment. Notons que pour qu'une chaîne de Markov puisse converger vers un état stationnaire, elle doit être irréductible et apériodique (Robert, 2001). L'algorithme atteint normalement cet état stationnaire si la chaîne respecte ces deux conditions. Ainsi, il est possible de trouver l'arbre optimal de

cette chaîne, appelé arbre MAP (maximum *a posteriori*), et l'arbre de consensus (l'arbre moyen).

4.2.2.1. *Comment proposer un nouvel arbre ?*

À l'étape 2 de l'algorithme Metropolis-Hastings, nous devons proposer un nouvel arbre, mais comment proposer celui-ci ? Il n'existe pas d'espace fixe d'arbres. Si un tel espace existait, il serait trop gros pour pouvoir l'analyser. Le graphe de l'espace d'arbres est plutôt un espace imagé. Mais, il n'existe pas réellement d'ordonnancement des arbres. Donc, il n'y a pas d'hiérarchie d'arbres. De plus, cet espace d'arbres imaginaire change à chaque fois que nous changeons le nombre d'espèces. Normalement, lorsque nous effectuons du Metropolis-Hastings pour proposer un nouvel arbre nous devons prendre un arbre voisin. Le mécanisme de transition est un mécanisme très important dans un algorithme MCMC. Nous devons pouvoir appliquer une transformation à l'état courant pour obtenir un état voisin qui ne soit ni trop proche, ni trop loin. Si celui-ci était trop proche, la chaîne progresserait trop lentement et prendrait trop de temps à converger vers la distribution stationnaire. Si par contre les propositions concernent des états trop distants, elles ont beaucoup moins de chance d'être acceptées et la convergence de la chaîne en sera influencé. Ainsi,

(1) Si R est près de 1, alors le nouvel arbre est souvent accepté ;

(2) Si R est égal à 1, alors le nouvel arbre est toujours accepté ;

(3) Si R est près de 0, alors nous rejettons le nouvel arbre, il est donc rarement accepté.

De cette manière, nous pouvons nous questionner sur l'importance de construire un mécanisme de transition pour trouver un arbre voisin dans un espace non ordonné.

4.2.2.2. *Méthodes de transition*

D'autres chercheurs ont également développé leur mécanisme de transition. Larget *et al.* (2005) présentent quatre mécanismes de transition différents qui permettent, selon eux, de parcourir l'espace d'états de manière efficace.

Voici les quatre mécanismes présentés par Larget *et al.* (2005).

Mécanisme 1 (voir les figures 4.2.1 et 4.2.2) :

(1) choisir un noeud interne au hasard (par exemple le noeud O) ;

(2) choisir au hasard deux des trois noeuds voisins (par exemple les noeuds A et C) ;

(3) choisir une nouvelle position entre les deux choisis à l'étape 2 (par exemple la position S_{31}) ;

(4) déplacer le noeud interne (de l'étape 1) à la nouvelle position.

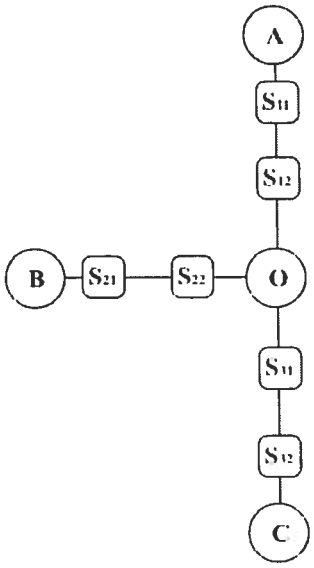


FIGURE 4.2.1. L'arbre de départ pour la mécanisme 1

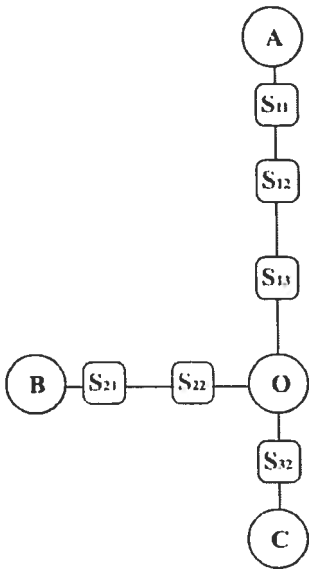


FIGURE 4.2.2. L'arbre final pour la mécanisme 1

Mécanisme 2 (voir les figures 4.2.3 et 4.2.4) :

(1) choisir une branche au hasard (par exemple la branche t1) ;

(2) choisir au hasard deux noeuds adjacents à la branche, un de chaque côté de la branche (soit les noeuds G et B) ;

(3) échanger les positions relatives des deux noeuds choisis à l'étape 2.

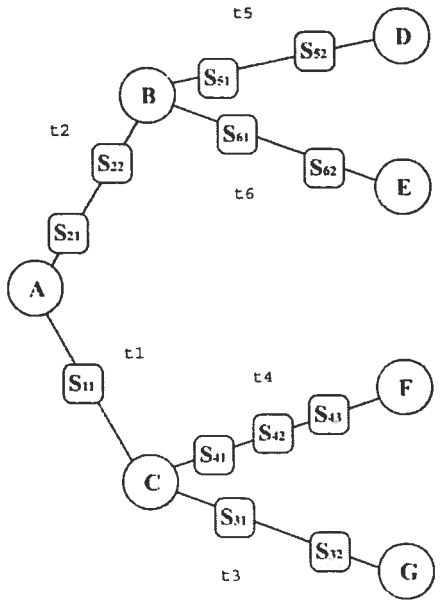


FIGURE 4.2.3. L'arbre de départ pour la mécanisme 2

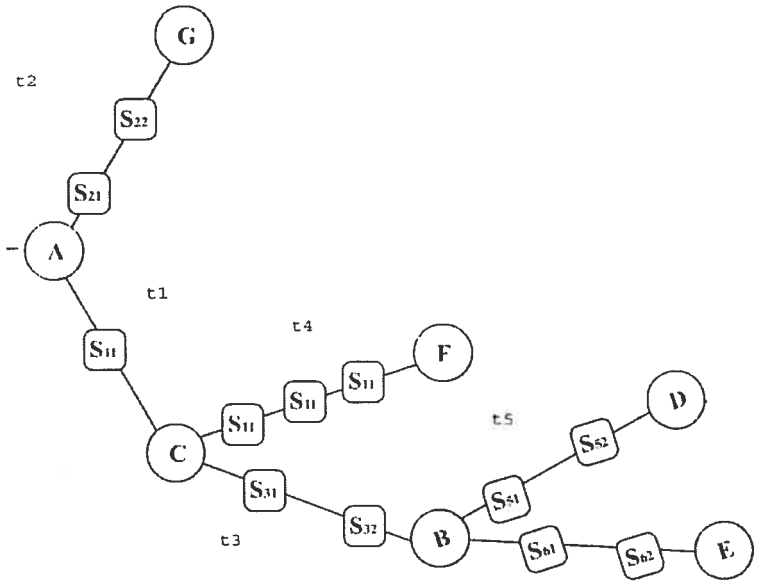


FIGURE 4.2.4. L'arbre final pour la mécanisme 2

Mécanisme 3 (voir les figures 4.2.5, 4.2.6 et 4.2.7) :

(1) choisir une branche interne au hasard (soit t1) ;

(2) éliminer la branche et les noeuds liés à cette branche (soit les noeuds A et C), ceci donne deux arbres ;

(3) choisir une branche dans chaque arbre (par exemple les branches t5 et t11) ;

(4) choisir une position sur chacune des deux branches (par exemple les positions S_{52} et S_{111}) ;

(5) créer deux nouveaux noeuds sur chacune des positions de l'étape 4 ;

(6) relier les deux noeuds en optant au hasard pour choisir l'ancêtre des deux.

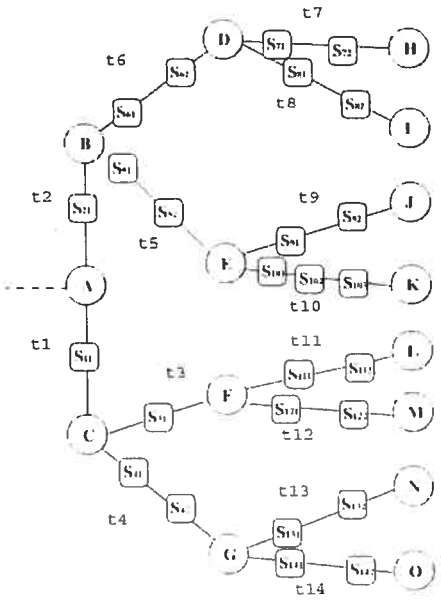


FIGURE 4.2.5. L'arbre de départ pour la mécanisme 3

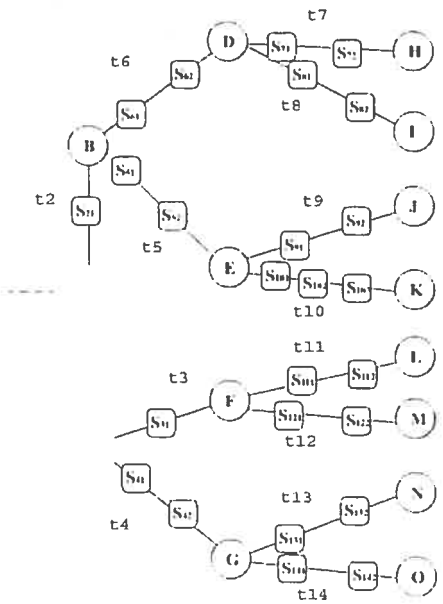


FIGURE 4.2.6. L'arbre intermédiaire pour la mécanisme 3

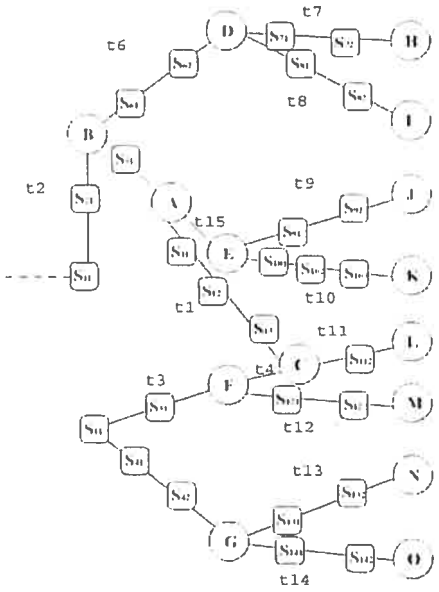


FIGURE 4.2.7. L'arbre final pour la mécanisme 3

Mécanisme 4 (voir les figures 4.2.8 et 4.2.9) :

(1) choisir deux espèces vivantes de l'étude (par exemple A et D) ;

(2) choisir un noeud interne sur le trajet qui relie les deux espèces de l'étape 1 (soit le noeud Z) ;

(3) choisir une nouvelle position qui est sur le chemin des deux espèces (soit la position S_{62}) ;

(4) prendre l'autre partie du noeud interne de l'étape 2 (soit la branche t4) et le déplacer à la position de l'étape 3.

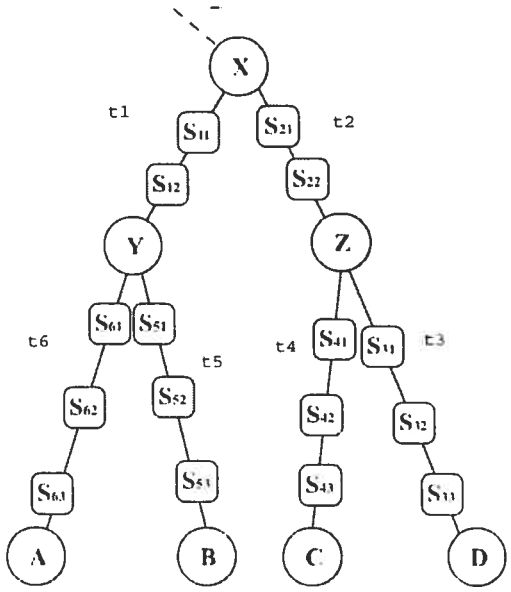


FIGURE 4.2.8. L'arbre de départ pour la mécanisme 4

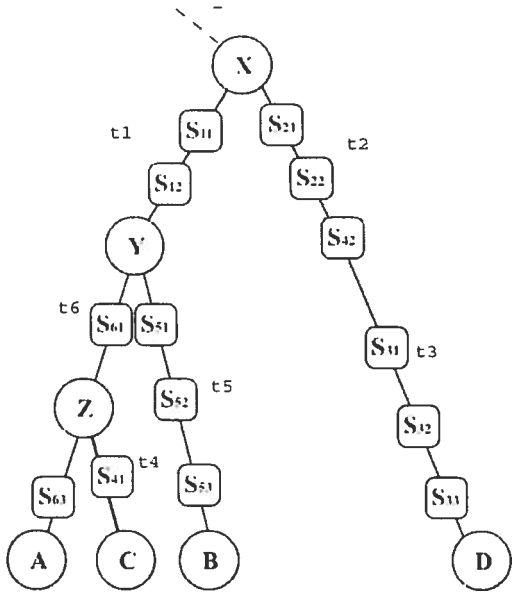


FIGURE 4.2.9. L'arbre final pour la mécanisme 4

Nous remarquons que le premier mécanisme change la longueur des branches, mais ne change pas la topologie de l'arbre, alors que les trois autres mécanismes le font. Ce qui implique que le premier mécanisme restera toujours sur la même topologie. Donc, il ne fait pas de transitions dans l'espace des arbres. Il est donc inutile de le considérer comme étant un processus de transitions. Toutefois, les trois autres mécanismes sont utiles pour effectuer le processus de transitions et sont les plus utilisés par les biologistes optant pour une approche bayésienne dans la reconstruction des arbres phylogénétiques.

Nous avons également décidé de développer une méthode de proposition d'arbre qui se base sur le nombre de noeuds des arbres :

(1) nous sélectionnons aléatoirement un nombre de noeuds à partir d'un poids qui est accordé par la proportion d'arbres possèdant ce nombre de noeuds sur le nombre d'arbres au total ;

(2) nous devons faire une permutation aléatoire des espèces ;

(3) nous construisons une distribution d'arbres associés au nombre de noeuds sélectionnés.

Cette méthode diffère des autres, puisqu'elle ne se base pas sur la transformation directe d'un arbre, mais plutôt d'une famille d'arbres (les arbres avec le même nombre de noeuds). De cette manière, elle visite l'espace des arbres plus rapidement, car elle ne demeure pas sur des arbres similaires à l'arbre de départ.

4.2.3. Maximum *a posteriori*

Dans l'inférence bayésienne, la méthode d'estimation du maximum *a posteriori*, dit MAP, est utilisée afin de trouver le maximum d'une fonction dans un contexte bayésien. Le MAP se distingue du maximum de vraisemblance par le fait qu'il est le maximum d'une distribution qui incorpore une distribution *a priori*.

Supposons que nous voulons estimer le paramètre θ pour maximiser la vraisemblance $L(\theta) = f(x|\theta)$ où x est un jeu de données incomplet. Soit f la distribution d'échantillonnage de x, donc, $f(x|\theta)$ est la fonction de x pour le paramètre θ. Alors $f(x|\theta)$, est considérée comme une fonction de θ, et $\hat{\theta}_{ML} = \arg\max_\theta f(x|\theta)$ est l'estimation du maximum de vraisemblance de θ.

Maintenant supposons que π est la distribution *a priori* pour θ. En se basant sur le théorème de Bayes, nous obtenons la distribution *a posteriori* :

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)d\theta}.$$

L'estimateur du maximum *a posteriori* cherche le mode de la distribution *a posteriori*, $\pi(\theta|x)$, en fonction de θ. Alors,

$$\begin{aligned}\hat{\theta}_{MAP}(x) &= \arg\max_\theta \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta} \\ &= \arg\max_\theta f(x|\theta)\pi(\theta).\end{aligned}$$

En utilisant l'estimateur MAP, nous allons obtenir l'arbre le plus probable à partir de l'approche bayésienne. Par conséquent en suivant la chaîne du MCMC, le MAP sera l'arbre qui représente le maximum de la distribution *a posteriori*. Toutefois, il existe d'autres manières d'obtenir le MAP, comme nous le verrons un peu plus loin, section 4.2.5 sur le ESE (Exploration, Sélection et Estimation).

4.2.4. Consensus

Tel que nous l'avons vu dans la sous-section précédente, la méthode MCMC va nous permettre d'obtenir un arbre optimal. De plus, elle donne une mesure de validation (support) pour n'importe quelle phylogénie. L'arbre de consensus donne une moyenne (probabilité) des regroupements des espèces qui sont revenus lors du rééchantillonnage du MCMC. Les branches étant présentées avec des pourcentages donnent les groupes monophylétiques (espèces partageant une affinité phylogénétique) qui se retrouvent le plus souvent ensemble, calculer à partir du MCMC (voir figure 4.2.10 et tableau 4.2.1). En contraste, la méthode du bootstrap et celle du jackknife utilisées pour valider les arbres obtenus en utilisant la méthode de maximum de vraisemblance donnent des mesures qui ne sont pas réellement des probabilités. En effet, il a été démontré (Paul-Michael Agapow, 2003) que les méthodes bootstrap et jackknife tendent vers des courbes sigmoïdales qui surestiment lorsque les valeurs sont élevées et qui sous-estiment pour des valeurs

TABLEAU 4.2.1. Nombre de répétitions des regroupements d'espèces dans le rééchantillonnage du MCMC (nombre d'itération = 1000)

Noeud	Espèces liées	nombre de répétitions du noeud
1	espèce A et espèce B	954
2	ancêtre 2 et ancêtre 3	901
3	ancêtre 1 et ancêtre 2	971

faibles. Ainsi, nous pouvons seulement dire qu'une valeur bootstrap (ou jackknife) de 75 % est meilleure que 74 % et qu'une valeur de 76 % est meilleure que 75 %.

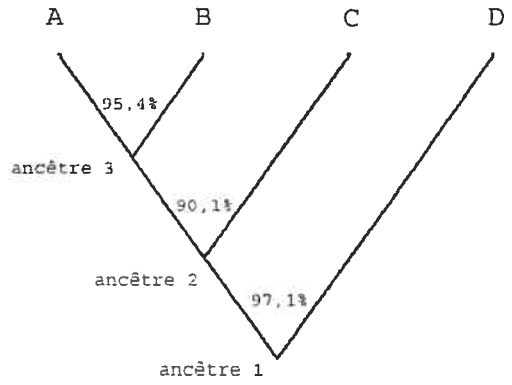


FIGURE 4.2.10. Un exemple d'arbre de consensus

4.2.5. ESE

Il existe d'autres algorithmes que celui de Metropolis-Hastings pour obtenir le MAP. La procédure du « Exploration, Sélection et Estimation », dit ESE (Destrempes *et al.*, 2005), est un algorithme très similaire à celui de Metropolis-Hasting. Toutefois, il diffère à l'étape 3 de l'algorithme (voir section 4.2.2). Au lieu de stagner à certains endroits comme peut le faire l'algorithme Metropolis-Hastings, le ESE se promène toujours dans l'espace des arbres. Ainsi pour ne pas rester pris, le ESE requiert le départ de plusieurs chaînes (soit deux au minimum)

en simultanée. En suivant un algorithme (voir la sous section 4.2.5.1) la méthode du ESE permet à une chaîne stagnante de faire une transition dans l'espace, en passant par la meilleure solution courante d'une autre chaîne. De cette manière, ces chaînes accepteront plus souvent les propositions de transition et se promèneront plus que les chaînes du Metropolis-Hastings. Toutefois, nous ne pouvons pas construire une mesure de consensus car la distribution est biaisée. Comme ces chaînes se promènent plus, elles peuvent obtenir des arbres plus probables que ceux obtenus par la méthode du Metropolis-Hastings.

4.2.5.1. *Algorithme ESE*

Soit f la fonction que nous cherchons à évaluer dans un espace fini A. Donc, dans notre cas, la fonction est la probabilité de la topologie de l'arbre et l'espace est l'espace des arbres. Soit ζ une structure de graphes connectés (voir la section 1.3) dans A. $N(T)$ le voisinage de $x \in A$, dans la structure de graphes ζ. Posons $\overrightarrow{T} = (T_1, T_2, T_3, ..., T_{m-1}, T_m)$, m arbres possibles où $m \geq 2$ et $\alpha(\overrightarrow{T})$ est la meilleure solution courante.

L'algorithme ESE se fait comme suit.

Tout d'abord, nous devons mentionner que l'algorithme ESE demande le départ de plusieurs chaînes en simultanée. Donc, nous expliquerons le déroulement pour une chaîne.

(1) Initialisation : Choisir aléatoirement un arbre dans la population initiale $\overrightarrow{T} = (T_1, T_2, T_3, ..., T_{m-1}, T_m)$.

(2) Exploration : L'algorithme doit se promener dans l'espace A à partir d'une proposition de transition, voir section 4.2.2.1.

(a) Ajuster $\alpha(\overrightarrow{T})$ en fonction de f à chaque itération.

(b) L'exploration et la sélection se font de la manière suivante, pour chaque itération, $i = 1, 2, 3, ..., m-1, m$, nous devons calculer deux paramètres qui nous permettent de choisir entre l'exploration et la sélection. Soit le paramètre $p = (i+2)^{\frac{-1}{C}}$ où C est une constante positive. Si p est inférieur à $u \sim Uniforme(0, 1)$, alors nous sommes dans le cas de l'exploration, ce qui signifie que l'on remplace T_i par $T_i' \in N(T_i)$. Si p est supérieur à u,

alors nous remplaçons T_i par $\alpha(\overrightarrow{T})$ (mais c'est ici que les autres chaînes interviennent, nous remplaçons T_i par la meilleure solution courante de toutes les chaînes, donc, $\alpha(\overrightarrow{T})$ de toutes les chaînes).

En pratique, $\alpha(\overrightarrow{T})$ estime le MAP de la fonction. Alors, la méthode ESE donne un estimé de l'arbre le plus probable.

Ainsi, en se promenant dans l'espace des arbres avec moins de restriction, comparativement à la méthode du Metropolis-Hastings, nous obtenons un maximum plus élevé. Cependant, comme nous obtenons un échantillon biaisé, il est impossible de construire un arbre de consensus avec le ESE.

4.2.6. Distribution *a priori*

Comme nous l'avons indiqué au début du chapitre, il est possible d'attribuer des probabilités *a priori* à certains arbres. Toutefois, il est difficile de faire suivre une densité *a priori* standard à ces arbres, puisqu'il n'existe pas d'ordonnancement pour les arbres.

Les paramètres tels que la longueur des branches peuvent être modélisés par une densité *a priori*. Incidemment, nous savons qu'en général une branche tend à être longue plutôt que courte. Puisque les mutations viables sont souvent très difficiles à obtenir, ce qui a pour conséquence d'allonger les branches, plus le temps entre 2 mutations s'écoule plus une branche sera longue. Notons qu'en général, seule une mutation sur trois entraînera un changement de la protéine finale. Ainsi, nous avons décidé de donner une distribution bêta modifié (β_m) de paramètres a et b à nos longueurs de branches, c'est-à-dire $\Pi(z) \propto e^{az}/(1+e^z)^{a+b-1}\mathbb{I}_{(0,\infty)}(z)$. Cette densité est obtenue à l'aide d'un changement de variables démontré dans le théorème suivant :

Théorème 4.2.1. *Si* $t \sim \beta(a,b)$, *alors* $z = \log\frac{t}{1-t} \sim \beta_m(a,b)$.

DÉMONSTRATION. Soit

$$t = \frac{e^z}{1+e^z}$$

$$t(1+e^z) = e^z$$

$$t = (1-t)e^z$$
$$e^z = \frac{t}{1-t}$$
$$z = \log \frac{t}{1-t}.$$

Donc, le changement de variables donne :

$$e^z dz = \frac{1-t+t}{1-t} dt$$
$$\frac{t}{1-t} dz = \frac{1}{1-t} dt$$
$$dz = \frac{1}{t} dt.$$

Ainsi, la densité de t est donnée par :

$$\begin{aligned}
\pi(z) &\propto \left(\frac{e^z}{1+e^z}\right)^{a-1} \left(1 - \frac{e^z}{1+e^z}\right)^{b-1} \left(\frac{e^z}{1+e^z}\right) \\
&= \left(\frac{e^z}{1+e^z}\right)^{a-} \left(1 - \frac{e^z}{1+e^z}\right)^{b-1} \\
&= e^{az}/(1+e^z)^{a+b-1} \\
&\sim \beta_m(a,b).
\end{aligned}$$

□

Donc, les longueurs de branches suivent une distribution bêta modifiée avec paramètres a et b ($z \sim \beta_m(a,b)$).

4.2.7. Convergence des chaînes

La convergence des chaînes est un autre aspect auquel nous devons nous préoccuper. Les chaînes de Markov peuvent souvent avoir tendance à rester « prises » à certains endroits et ne visitent donc pas l'espace des arbres dans son ensemble. Pour éviter ce genre de situation, l'utilisation d'une technique nommée

les couples de Metropolis est souvent employée, notée MCMCMC. Cette technique prévoit l'utilisation de plusieurs chaînes en parallèle. Les chaînes partent à différents points de départ, donc à différentes topologies. Il existe deux types de chaînes. Le premier type de chaîne est celui qui converge vers la distribution *a posteriori* et qui se promène moins rapidement. L'autre type de chaîne est celui dont les probabilités d'acception sont modifiées afin d'accroître le taux de visite des différents états.

De cette manière, ces chaînes accepteront plus souvent les propositions de transition et se promènent plus que les autres. Toutefois, nous ne pouvons pas les échantillonner car la distribution qui en résulte est biaisée. Ces chaînes sont utilisées dans le but d'explorer l'espace des états. En fait, ceci permet au type de chaîne qui bouge moins d'échanger de temps en temps de position avec une des chaînes qui bougent plus, ceci lui évitant de rester prise à un endroit. Incidemment, l'utilisation de cet algorithme MCMCMC, permet de s'assurer que la chaîne convergera, et diminuera le nombre de simulations à faire. Par contre, le temps de simulation sera plus élevé à cause du nombre de chaînes que nous simulons en même temps.

4.3. Comparaison des méthodes phylogénétiques

Cette dernière section du quatrième chapitre explique les méthodes employées afin de comparer les arbres. Comment savoir si deux arbres sont significativement différents ou pas ? Plusieurs méthodes ont été développées par les biologistes et les mathématiciens. Dans le cadre de cette étude, nous allons employer quatre méthodes de comparaison :

(1) le test des sites gagnants (Templeton, 1983 ; Wilson, 1988 ; Felsenstein, 1984) ;

(2) le test des rangs signés de Wilcoxon (Templeton, 1983) ;

(3) test-t à échantillons appariés (Swofford *et al.*, 1996a) ;

(4) matrice des distances dans un arbre phylogénétique probabiliste (voir section 4.3.4).

Les arbres I (voir figure 4.3.1) et II (voir figure 4.3.2) seront nos arbres de référence pour expliquer la méthodologie qu'emploie chaque méthode. Le tableau 4.3.1 présente le logarithme des vraisemblances aux différents sites pour nos deux arbres.

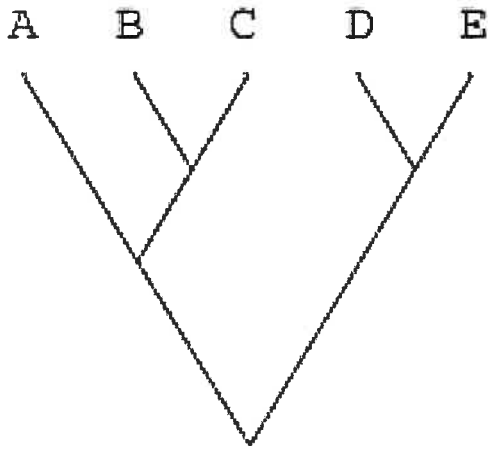


FIGURE 4.3.1. L'arbre I

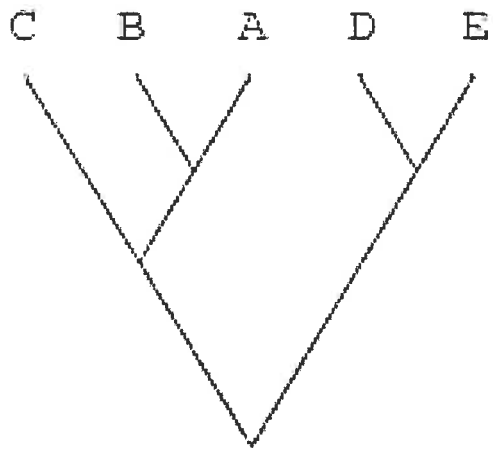


FIGURE 4.3.2. L'arbre II

4.3.1. Le test des sites gagnants

Pour chaque site, nous devons calculer quel arbre a le meilleur résultat, c'est-à-dire la plus grande probabilité pour le site en question. Donc, nous calculons le nombre de fois que chaque arbre est gagnant pour chaque site. Pour vérifier la différence entre les deux arbres, nous comparons par rapport à une distribution

TABLEAU 4.3.1. Tableau du logarithme des vraisemblances aux différents sites pour les arbres I et II

arbre \ site	1	2	3	4	5	6	7	8	$\log L$
I	-2,971	-4,483	-5,673	-5,883	-2,691	-8,003	-2,971	-2,691	-35,366
II	-2,983	-4,494	-5,685	-5,898	-2,700	-7,572	-2,987	-2,705	-35,024
différence	0,012	0,011	0,012	0,015	0,009	-0,431	0,016	0,014	-0,342
rang	3,5	2	3,5	6	1	8	7	5	
signe	+	+	+	+	+	-	+	+	

binomiale les résultats obtenus et s'ils diffèrent significativement de $1/2$, alors nous pouvons conclure que les arbres sont différents.

En utilisant la méthode du test des sites gagnants pour notre exemple, voir le tableau 4.3.1, nous obtenons les résultats suivants : l'arbre I possède une plus grande probabilité pour tous les sites, sauf le site 6. Donc, sept sites sont gagnants par l'arbre I et un seul site par l'arbre II. Nous avons une observation 7 ($o = 7$) pour l'arbre I, si les deux arbres étaient similaires, nous aurions eu 4 sites gagnants par l'arbre I et 4 par l'arbre II. Une telle comparaison se mesure par une statistique khi-deux (χ^2). Le nombre de fois que l'arbre I possède une plus grande vraisemblance que l'arbre II suit une loi binomiale : $x \sim Binomiale(n, p)$ où n est le nombre de sites et p est $\frac{1}{2}$, sous l'hypothèse que les deux arbres soient similaires. L'espérance de x est égale à np et sa variance vaut $np(1\text{-}p)$.

Alors pour notre exemple, la statistique khi-deux vaut :

$$\begin{aligned}\chi_1^2 &= \frac{(o-np)^2}{np(1-p)} \\ &= \frac{(7-4)^2}{2} \\ &= 4{,}5.\end{aligned}$$

Nous avons une valeur-p =0,0339. Par conséquent, nous rejetons l'hypothèse nulle au niveau 5%. Donc, les deux arbres sont significativement différents selon la méthode du test des sites gagnants.

4.3.2. Le test des rangs signés de Wilcoxon

La méthode du test des rangs signés de Wilcoxon est employée dans des contextes statistiques non paramétrique. Elle s'applique en phylogénétique depuis un certain nombre d'années. Cette méthode consiste à calculer sur chaque site la différence des vraisemblances entre deux arbres. Ainsi, nous obtenons un vecteur de différences entre les arbres, d. Ensuite, nous devons mettre en ordre croissant les valeurs absolues des différences pour pouvoir sommer les rangs des différences positives. Si la grandeur de l'échantillon est plus grande que 16 sites (ce qui est généralement le cas), alors le test peut se faire par une approximation normale.

$$E(W_+) = \frac{1}{4}n(n+1),$$
$$\sqrt{Var(W_+)} = \sqrt{\frac{n(n+1(2n+1)}{24}},$$

où $E(W_+)$ représente l'espérance de la somme des rangs positifs, $\sqrt{Var(W_+)}$ donne l'écart type de cette somme et n le nombre de sites dans le jeu de données.

Nous comparons la somme des rangs à une distribution normale suivant les paramètres $E(W_+)$ et $Var(W_+)$, ainsi nous pouvons vérifier s'il existe une différence significative entre les deux arbres (Rice, 1995).

À des fins pédagogiques, nous allons supposer que notre échantillon est plus grand que 16 sites, ainsi, nous pourrons utiliser l'approximation par une loi normale de la statistique de Wilcoxon. Nous obtenons une seule différence négative (site 6), voir tableau 4.3.1 (W_+ est le minimum des signes, ce qui signifie que comme nous avons 1 signe négatif et sept signes positifs, nous allons prendre le signe négatif). C'est la plus grande des différences, alors notre statistique W_+ vaut 8. En normalisant, nous obtenons les résultats suivants :

$$E(W_+) = \frac{1}{4}n(n+1) = \frac{1}{4}8(8+1)$$
$$= 18,$$

$$\begin{aligned}\sqrt{Var(W_+)} &= \sqrt{\frac{n(n+1(2n+1)}{24}} \\ &= \sqrt{\frac{8(8+1(2(8)+1)}{24}} \\ &= 7{,}14143.\end{aligned}$$

Alors,

$$\begin{aligned}z &= \frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}} \\ &= \frac{8-18}{7{,}14143} \\ &\cong -1{,}40.\end{aligned}$$

Notre valeur-p est égale à 0,08076. Ainsi, au niveau $\alpha = 0{,}05$, nous ne rejetons pas l'hypothèse de l'égalité des arbres. Alors, nos deux arbres ne sont pas significativement différents selon le test des rangs signés de Wilcoxon.

4.3.3. Test-t à échantillons appariés

Un test-t est un test d'hypothèses statistique dans lequel la statistique du test suit une distribution Student si l'hypothèse nulle est vraie. L'hypothèse nulle est que les deux arbres sont équivalents. Dans notre contexte, nous supposons donc que la différence de la vraisemblance entre deux arbres sur un même site, suit une distribution de Student. Ainsi, notre test d'hypothèses est en réalité un t-test à échantillons appariés. Ce test permet d'observer la différence des vraisemblances entre deux arbres. Encore une fois, nous comparons les vraisemblances sur chacun des sites. Ces différences appariées donnent une moyenne, $\bar{d}$, et une variance, s_d sur les différences. Alors, nous obtenons le test d'hypothèses suivant :

Si nous voulons confronter

$$H_0 : \delta = 0 \text{ et } H_1 : \delta \neq 0.$$

La différence des vraisemblances sur un site est représentée par $d_i = x_i - y_i$ (δ est la différence moyenne sur un site au niveau de la population), où x_i est la

vraisemblance du premier arbre au site i et y_i est la vraisemblance du deuxième arbre au site i. Alors,

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i,$$

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2},$$

$$t_{n-1} = \frac{\sqrt{n} \times \bar{d}}{s_d},$$

où n est le nombre de sites. Ainsi, si nous rejetons l'hypothèse nulle, alors ceci signifie qu'il existe une différence significative entre les deux arbres.

En ce qui concerne notre exemple, le test-t pour échantillons appariés donne les résultats suivants :

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{8}\sum_{i=1}^{8} d_i = -0{,}04275,$$

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2} = 0{,}15689282,$$

$$t_7 = \frac{\sqrt{8} \times -0{,}04275}{0{,}15689282} = -0{,}77075.$$

Notre valeur-p est égale à 0,2779, donc nous ne pouvons pas rejeter l'hypothèse nulle au niveau $\alpha = 0{,}05$, que la moyenne des différences entre les fonctions de vraisemblance des deux arbres est égale à 0. Alors, il n'existe pas de différence significative entre les deux arbres selon la méthode du test-t pour échantillons appariés.

4.3.4. Matrice des distances dans l'arbre

Nous avons développé cette dernière méthode afin de répondre à un besoin de comparer deux arbres de manière topologique. Dans le contexte statistique de la phylogénie, la grande majorité des méthodes de comparaison utilisent les vraisemblances sur les sites. Dans le contexte de parcimonie, la comparaison des

partitions des espèces est la méthode utilisée pour comparer les arbres. Nous voulions créer une méthode qui confronte les topologies, mais non par partitions, puisque deux arbres qui ont les mêmes partitions peuvent être très différents (voir figure 3.3.1). Donc, nous avons développé une méthode qui prend en compte la topologie, mais également la longueur des branches d'un arbre. La matrice des distances dans l'arbre permet de comparer deux arbres de manière topologique, mais sans avoir le problème des différentes longueurs de branches. Cette méthode ne nous dit pas si deux arbres sont significativement différents, mais nous donne un indice sur la différence. Ainsi, cet indice est élevé si la différence entre les arbres est grande. Elle nous donne l'opportunité de comparer des méthodes statistiques, en occurrence l'analyse bayésienne et l'analyse par maximum de vraisemblance.

Le processus de la méthode est fort simple, nous créons une matrice de distances entre chaque espèce, noté D. Pour un arbre donné, nous calculons la distance minimale entre chaque paire d'espèces. La distance utilisée correspond à la somme des longueurs des branches. Dans le contexte où nous aurions quatre espèces dans notre arbre, la matrice D prendrait la forme suivante :

$$D = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{31} & d_{32} & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{pmatrix} \tag{4.3.1}$$

où d_{ij} est la distance qui sépare l'espèce i de l'espèce j. Comme la matrice est symétrique, alors nous obtenons l'égalité suivante :

$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{12} & 0 & d_{23} & d_{24} \\ d_{13} & d_{23} & 0 & d_{34} \\ d_{14} & d_{24} & d_{34} & 0 \end{pmatrix} \tag{4.3.2}$$

avec $d_{ij} = \sum_k t_k$, si k appartient au trajet entre i et j et t_k est la longueur de la branche k.

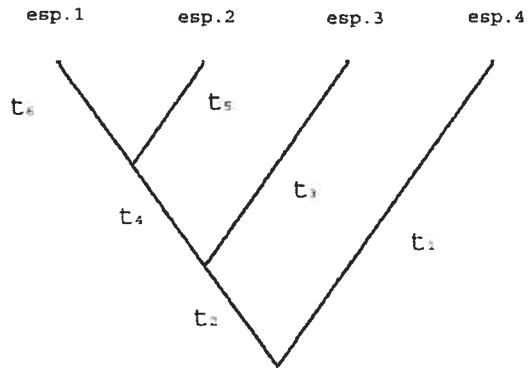


FIGURE 4.3.3. Arbre phylogénétique et les branches t_k

Pour expliquer le fonctionnement de cette méthode, nous allons prendre l'exemple des figures 4.3.3, 4.3.4 et 4.3.5. Prenons l'exemple de l'arbre original, il existe quatre branches entre l'espèce 1 et l'espèce 4. Alors,

$$\begin{aligned} d_{(1)14} &= t_1 + t_2 + t_4 + t_6 \\ &= 7 + 3 + 2 + 1{,}5 = 13{,}5. \end{aligned}$$

En continuant ce processus, nous obtenons les matrices $D_{(1)}$, pour l'arbre original (voir la matrice (4.3.3) et la figure 4.3.4) et $D_{(2)}$, pour l'arbre reconstruit (voir la matrice (4.3.4) et la figure 4.3.5).

$$D_{(1)} = \begin{pmatrix} 0 & 4 & 7{,}5 & 13{,}5 \\ 4 & 0 & 8{,}5 & 14{,}5 \\ 7{,}5 & 8{,}5 & 0 & 14 \\ 13{,}5 & 14{,}5 & 14 & 0 \end{pmatrix}, \tag{4.3.3}$$

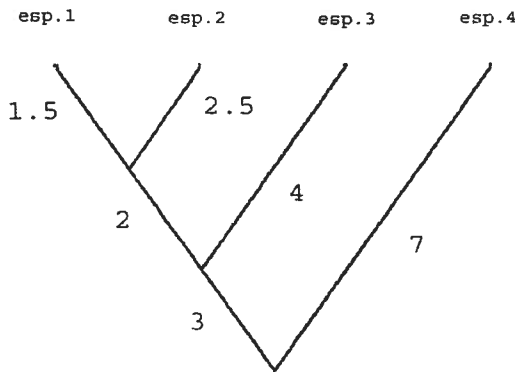


FIGURE 4.3.4. Arbre phylogénétique avec des longueurs de branches originales

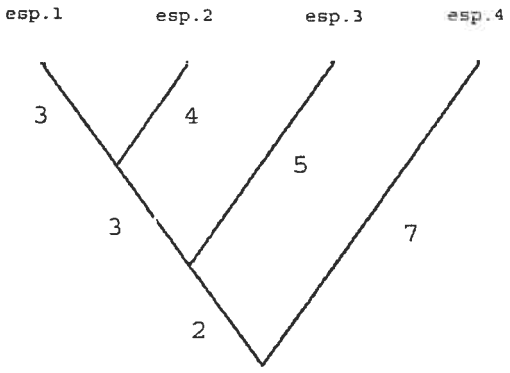


FIGURE 4.3.5. Arbre phylogénétique avec des longueurs de branches reconstruites

$$D_{(2)} = \begin{pmatrix} 0 & 7 & 11 & 15 \\ 7 & 0 & 12 & 16 \\ 11 & 12 & 0 & 14 \\ 15 & 16 & 14 & 0 \end{pmatrix}. \tag{4.3.4}$$

La distance entre deux arbres peut ainsi se mesurer par la différence des éléments de la partie triangulaire supérieure ou triangulaire inférieure des matrices D_1 et D_2, ou encore en divisant par deux la somme des différences des deux matrices.

$$\begin{aligned} d_{totale} &= \sum_{i,j|i>j} |d_{(1)ij} - d_{(2)ij}| \\ &= \sum_{i,j|i<j} |d_{(1)ij} - d_{(2)ij}| \\ &= \frac{1}{2} \sum_{i,j} |d_{(1)ij} - d_{(2)ij}| \\ &= 13. \end{aligned}$$

Cette distance permettra de comparer les méthodes d'analyse phylogénétique (voir la section 5.3).

4.4. Conclusion du chapitre

Depuis environ cinq ans, le monde de la phylogénétique commence à s'intéresser à la théorie bayésienne ce qui donne une autre perspective et un nouvel outil à une science qui demeure encore abstraite. Dans le contexte de la phylogénétique, la théorie du maximum de vraisemblance et l'approche bayésienne sont très similaires, car les deux emploient un calcul probabiliste pour trouver l'arbre phylogénétique. Toutefois, il demeure une différence entre ces deux méthodes qui réside dans le calcul de la probabilité de l'arbre. Le maximum de vraisemblance va trouver l'arbre le plus probable d'expliquer les séquences observées $P(D|T)$ (la probabilité des données moléculaires, D, sachant l'arbre phylogénétique, T). Tandis que l'approche bayésienne donnera l'arbre, ou un consensus d'arbres, qui est le plus probable d'être expliqué par les séquences observées $P(H|D)$ (la probabilité des hypothèses phylogénétiques, T, sachant les données moléculaires, D). Cependant, ce problème qui peut sembler anodin, engendre une toute nouvelle méthode d'analyse phylogénétique possédant énormément d'avantages, tels que

le fait de donner un échantillon. Donc, nous pouvons ainsi donner un intervalle de crédibilité à notre arbre, ce que la méthode du maximum de vraisemblance ne peut faire, puisqu'elle suit un processus heuristique pour trouver l'arbre le plus probable. De plus, l'approche bayésienne donne une vraie mesure de validation, arbre de consensus, contrairement à la méthode du maximum de vraisemblance qui approxime avec les méthodes du bootstrap et du jacknife. Ainsi dans le prochain chapitre, nous allons comparer, avec des jeux de données simulés et un vrai jeu de données, les deux méthodes en nous servant des tests de comparaison vus à la fin de ce chapitre.

Chapitre 5

SIMULATIONS, RÉSULTATS ET COMPARAISON DES APPROCHES PHYLOGÉNÉTIQUES

5.1. SIMULATIONS

Dans ce dernier chapitre, nous allons appliquer les théories développées dans les quatre premiers chapitres. La méthode par maximum de vraisemblance et l'approche bayésienne seront comparées sous deux aspects. En premier lieu, nous confronterons les deux méthodes par l'entremise d'une reconstruction des jeux de données obtenus à partir des modèles simulés. De cette manière, nous pouvons revalider les approches à partir de ces jeux de données. En deuxième lieu, nous utiliserons un vrai jeu de données pour pouvoir comparer les deux méthodes. Ainsi, il sera possible de constater les différences et illustrer les avantages de l'approche bayésienne vis-à-vis la méthode du maximum de vraisemblance. En conséquence, nous illustrons par implication (vrai jeu de données) et par la contraposée (reconstruction des jeux de données) que l'approche bayésienne est supérieure à la méthode du maximum de vraisemblance qui est considérée comme étant la norme en phylogénétique (méthode la plus employée). Ceci étant dit, pour l'approche bayésienne nous avons utilisé notre méthode de proposition d'arbres (voir section 4.2.2.2). Il existe d'autres mécanismes de proposition d'arbres. Par exemple, le programme Mr.Bayes (disponible sur le site : http ://mrbayes.csit.fsu.edu/)

emploie les mécanismes présentés par Larget *et al.* (2005) (voir section 4.2.2.2). Toutefois, nous voulions introduire une alternative à ce programme.

5.2. Reconstructions de jeux de données

La première étape de comparaison est la reconstruction de jeux de données. Nous allons valider les méthodes en allant par l'inverse. Le concept de la reconstruction de jeux de données est le suivant (voir la figure 5.2.1) :

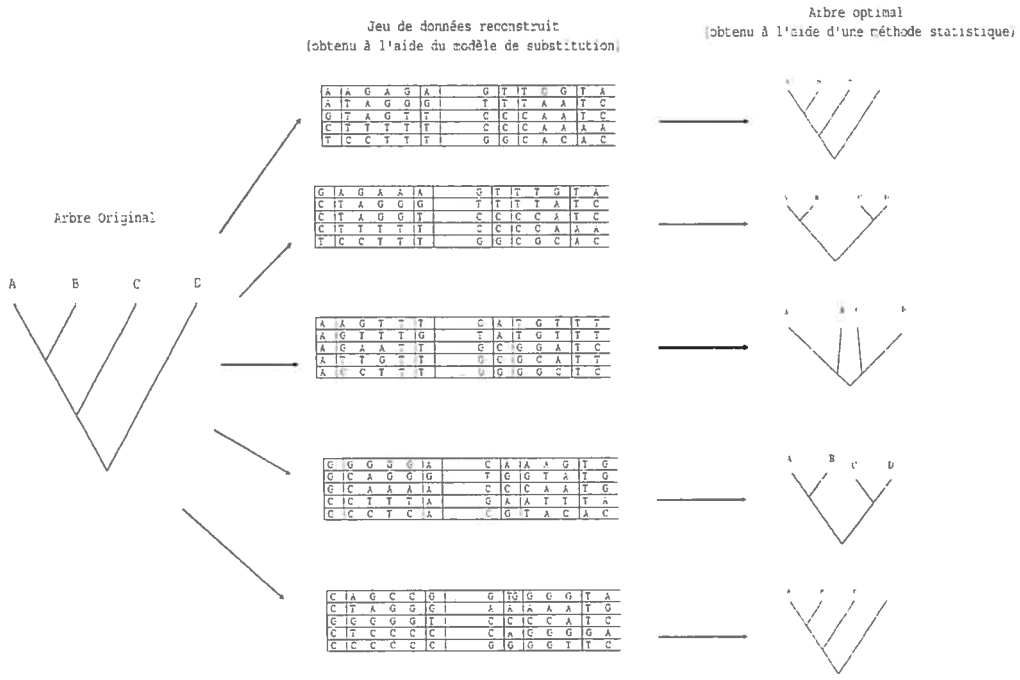


Figure 5.2.1. Le processus de la reconstruction de jeux de données

(1) construisons un arbre quelconque, noté T (appelé l'arbre original) ;

(2) reconstruisons un jeu de données, D, à partir de l'arbre et du modèle de substitution utilisé ;

(3) à partir du jeu de données, D, trouvons l'arbre optimal pour chacune des deux approches (T_{mle} : arbre obtenu par maximum de vraisemblance ; T_{bay} : arbre obtenu par approche bayésienne) ;

(4) comparer (à l'aide des tests vus à la section 4.3) les arbres obtenus par les deux approches, T_{mle} ou T_{bay}, à T, l'arbre original.

Dans le cadre de cette recherche, nous reconstruirons les jeux de données à partir du modèle de substitution Jukes-Cantor, JC (voir la section 3.2.1). Voici le déroulement de la fonction qui nous permettra d'obtenir les jeux de données.

Pour chaque reconstruction, nous devons spécifier le nombre de sites que nous voulons simuler. Ensuite, nous partons de l'ancêtre en lui donnant aléatoirement l'une des quatre nucléotides possibles (A, C, G ou T). À partir de l'ancêtre, nous pourrons simuler les nucléotides des descendants de l'ancêtre et finalement des vraies espèces du jeu de données. Pour ce faire, nous allons calculer la probabilité de mutation sur chaque branche (la probabilité variant entre 0 et 1) et nous allons comparer cette dernière à une valeur aléatoire générée à partir d'une distribution uniforme variant entre 0 et 1, notée Z. Si cette valeur aléatoire est supérieure à la probabilité de mutation, alors nous allons attribuer au descendant de l'ancêtre un autre nucléotide que celle de l'ancêtre. Comme nous utilisons le modèle de substitution de Jukes-Cantor, nous choisirons aléatoirement l'un des trois nucléotides différents de celui de l'ancêtre. Si Z est inférieure à la probabilité de mutation, nous attribuerons au descendant de l'ancêtre le même nucléotide que celui de l'ancêtre (voir figure 5.2.2).

Ainsi de suite, en simulant pour chacun des descendants et des espèces, nous pourrons compléter le jeu de données pour tous les sites. En répétant ces mêmes étapes à chaque site, nous obtenons un jeu de données (voir figure 5.2.3).

Ensuite, il ne reste qu'à trouver l'arbre optimal selon chacune des méthodes statistiques employées en phylogénétique (maximum de vraisemblance et approche bayésienne) et comparer ces arbres à l'arbre original (voir section 4.3).

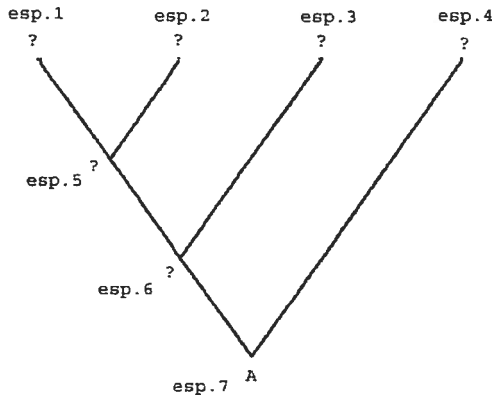


FIGURE 5.2.2. La reconstruction d'un jeu de données

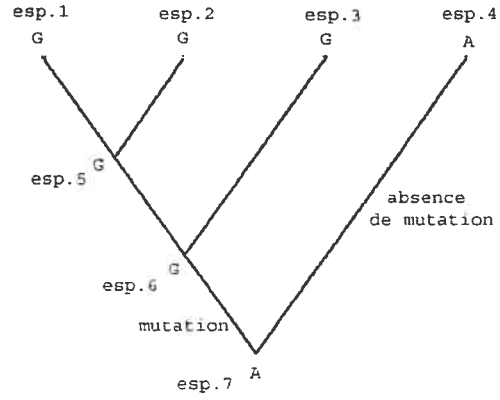


FIGURE 5.2.3. La reconstruction complète d'un jeu de données

5.2.1. L'arbre original

Comment sélectionner l'arbre original ? Dans le domaine de la phylogénétique, les arbres aux longues branches sont les plus problématiques (voir section 1.5). Les arbres aux longues branches sont des arbres :

(1) possédant de courtes branches entre les espèces et leurs descendants, et de longues branches entre ces descendants et leurs ancêtres (voir figure 5.2.4) ;

(2) possédant de longues branches entre les espèces et leurs descendants, et de courtes branches entre ces descendants et leurs ancêtres (voir figure 5.2.5).

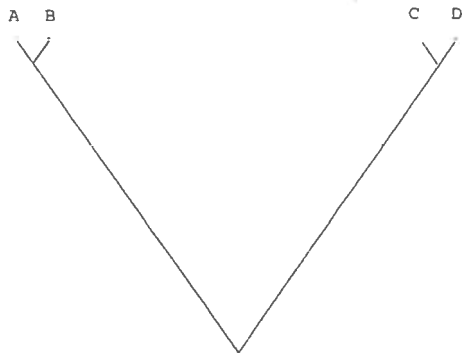


FIGURE 5.2.4. Arbre à courtes branches entre les espèces

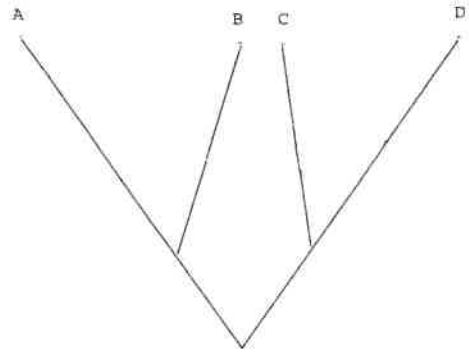


FIGURE 5.2.5. Arbre à longues branches entre les espèces

Incidemment, nous allons reconstruire pour ces deux arbres énumérés précédement. Dans la perspective de compléter l'étude, nous allons également resimuler pour un arbre typique, normal, aux longueurs de branche égales (voir figure 5.2.6).

5.2.2. Nombre de répétitions

Pour chaque arbre original, nous devons produire un nombre de répétitions du jeu de données. Les recherches en phylogénétiques qui font de la reconstruction de jeu de données construisent en moyenne 1000 répétitions du jeu de données. Nous pourrions faire un plus grand nombre de répétitions, toutefois, le temps de simulation augmenterait significativement (nous parlons ici de journées).

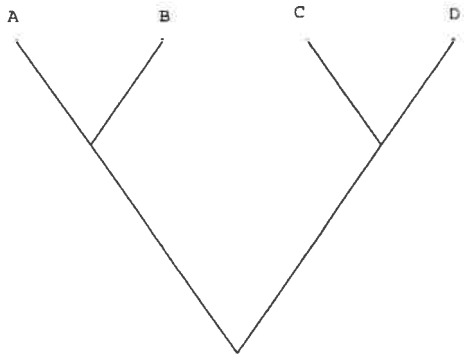


FIGURE 5.2.6. Arbre à longueurs de branche égales

5.2.3. Nombre d'espèces dans le jeu de données reconstruit

Pour pouvoir reconstruire un jeu de données, nous devons prédéfinir le nombre d'espèces que nous voulons avoir dans le jeu de données. Toutefois, il n'existe pas de norme, dans la reconstruction de jeu de données. Certaines recherches emploient quatre espèces et d'autre quinze espèces. Dans le cadre de cette étude, nous allons faire les reconstructions avec cinq et six espèces, en raison du temps de simulation. Remarquons que plus nous augmentons le nombre d'espèces plus le temps de simulation augmente (ici l'augmentation croît de manière exponentielle, alors, nous passons d'une simulation qui prend une journée à une simulation qui peut prendre des semaines), puisque pour chaque arbre original nous devons produire 1000 répétitions.

5.2.4. Nombre de sites dans le jeu de données reconstruit

Il ne reste qu'à déterminer le nombre de sites pour pouvoir entreprendre les simulations. Pour le nombre de sites, la plupart des recherches phylogénétiques utilisent un minimum d'au moins 500 sites pour pouvoir obtenir des arbres plausibles. Pour cause d'économie de temps de simulation, nous allons employer 1000 sites pour faire nos reconstructions de jeux de données.

TABLEAU 5.3.1. Description des simulations

Simulation	Arbre original	Nombre d'espèces	Figure
1	Arbre aux longueurs de branche égales	5	5.3.1
2	Arbre aux longueurs de branche égales	6	5.3.2
3	Arbre aux courtes branches entre espèces	5	5.3.3
4	Arbre aux courtes branches entre espèces	6	5.3.4
5	Arbre aux longues branches entre espèces	5	5.3.5
6	Arbre aux longues branches entre espèces	6	5.3.6

Ainsi, après avoir prédéterminé nos arbres originaux, le nombre de répétitions, le nombre d'espèces et le nombre de sites dans le jeu de données reconstruit, nous pouvons entreprendre les reconstructions des jeux de données.

5.3. RÉSULTATS DES RECONSTRUCTIONS DES JEUX DE DONNÉES

Cette section nous permettra d'observer les résultats obtenus à l'aide des deux approches (maximum de vraisemblance et approche bayésienne) pour les six reconstructions (simulations) de jeu de données. Ceci est présenté dans le tableau 5.3.1.

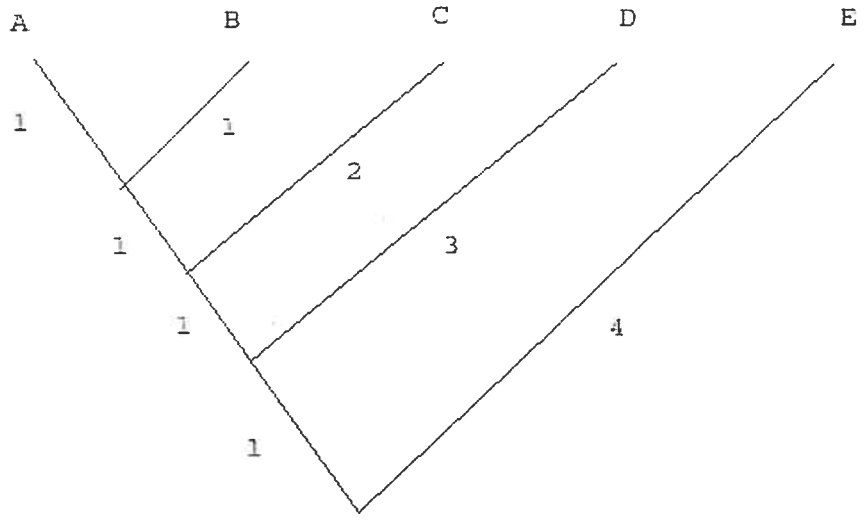


FIGURE 5.3.1. Arbre de la simulation 1

Pour chaque simulation, nous avons construit un tableau qui fait la comparaison entre l'arbre original et les 1000 arbres optimaux obtenus par la méthode du

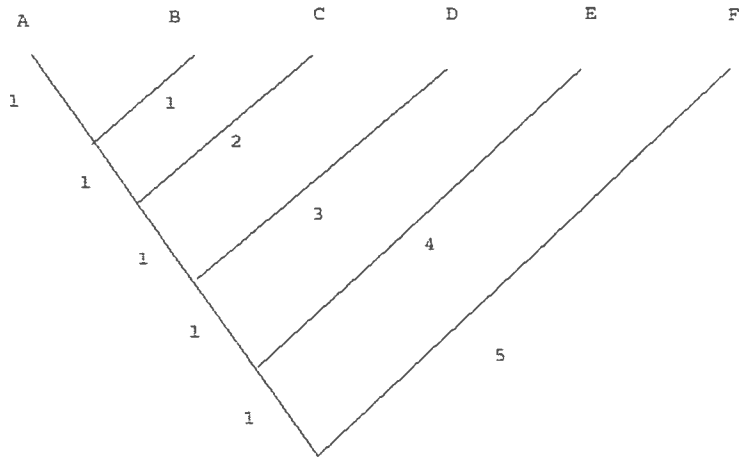


FIGURE 5.3.2. Arbre de la simulation 2

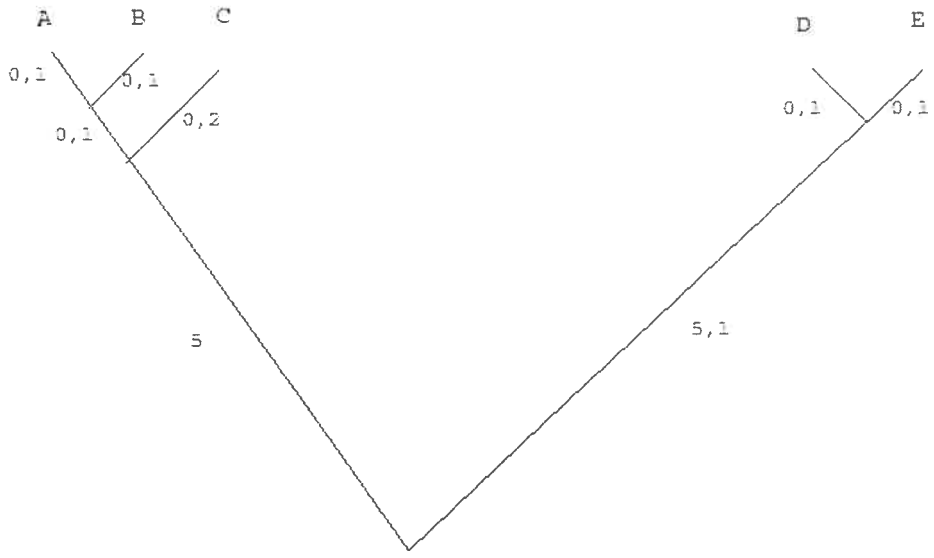


FIGURE 5.3.3. Arbre de la simulation 3

maximum de vraisemblance et la comparaison entre l'arbre original et les 1000 arbres optimaux obtenus par l'approche bayésienne.

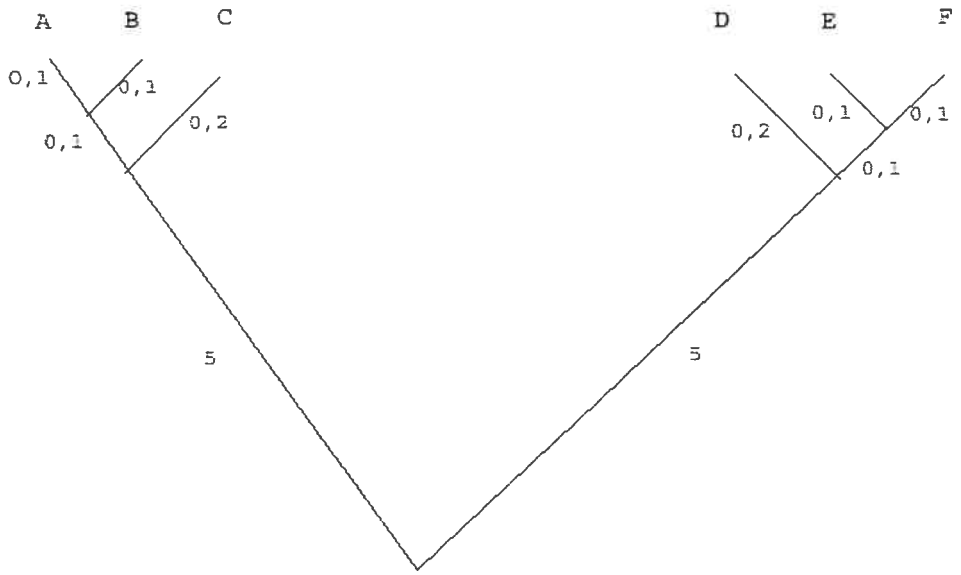


FIGURE 5.3.4. Arbre de la simulation 4

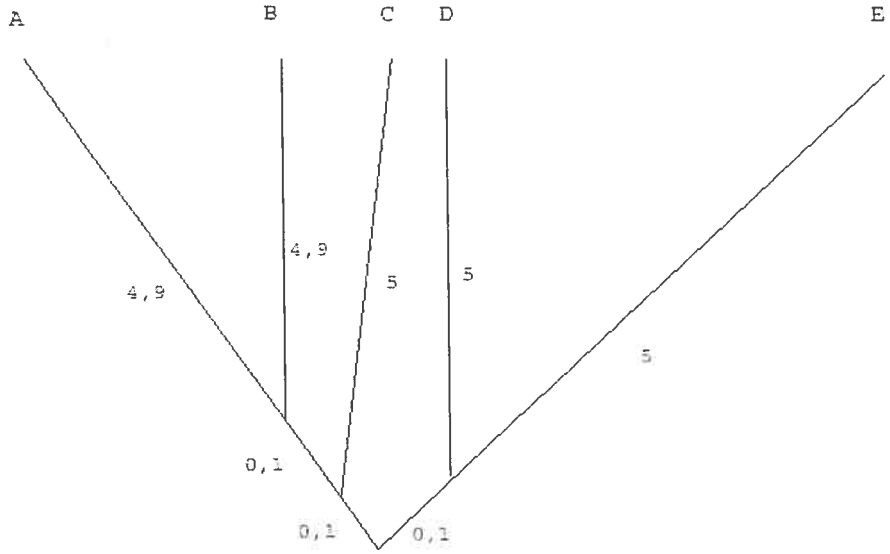


FIGURE 5.3.5. Arbre de la simulation 5

5.3.1. Première simulation

De cette manière pour la première simulation, nous avons cinq espèces séparées par des distances de branches égales (voir figure 5.3.1). Le tableau 5.3.2 nous donne un aperçu des résultats obtenus pour les approches de maximum de vraisemblance et bayésienne. Nous pouvons voir par l'entremise de ce tableau

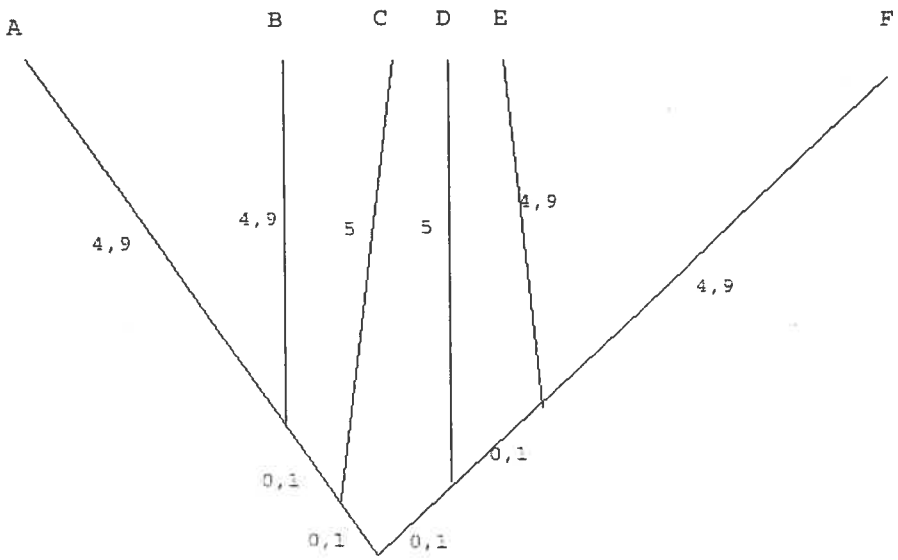


FIGURE 5.3.6. Arbre de la simulation 6

que l'approche bayésienne obtient de meilleurs résultats pour tous les tests. Selon le test des sites gagnants, l'approche bayésienne reconstruit mal (significativement différent) 271 arbres (arbre optimal de 271 jeux de données), tandis que la méthode du maximum de vraisemblance en reconstruit 335 arbres différents de l'arbre original. Pour le test des rangs signés de Wilcoxon et le test-t pour échantillons appariés, l'approche bayésienne construit différement 97 et 30 arbres respectivement, alors que la méthode du maximum de vraisemblance compte 159 et 35 arbres différents. Finalement le dernier test de matrice des distances, nous permet de constater que la distance entre les arbres reconstruits et l'arbre original est en moyenne de 6,937 pour l'approche bayésienne et de 30,220 pour la méthode par maximum de vraisemblance.

Une autre forme de comparaison que nous utilisons est celle des comparaisons directes, ce qui signifie, de comparer pour chaque jeu de données reconstruit quelle méthode est celle qui donne l'arbre le plus près. Donc, laquelle produit l'arbre le plus près (celle qui possède une distance inférieure) pour chacune des 1000 répétitions du jeu de données. Dans ce contexte, l'approche bayésienne l'emporte 691 fois contre 309 pour la méthode du maximum de vraisemblance.

TABLEAU 5.3.2. Résultats de la première simulation pour la méthode du maximum de vraisemblance (EMV) et l'approche bayésienne en fonction des tests de comparaison, à α =5 % (voir section 4.3)

Méthode \ Test	Sites gagnants nombre d'arbres rejetés	Rangs signés de Wilcoxon nombre d'arbres rejetés	t nombre d'arbres rejetés	Matrice des distances distance moyenne
EMV	335	159	35	30,220
Bayes	271	97	30	6,937

5.3.2. Deuxième simulation

Pour la deuxième simulation, nous avons procédé de la même manière que pour la première simulation, mais cette fois-ci avec six espèces séparées par des distances de branches égales (voir figure 5.3.2). Le tableau 5.3.3 nous donne un aperçu des résultats obtenus pour les approches de maximum de vraisemblance et bayésienne. Encore une fois, nous pouvons voir par l'entremise du tableau que l'approche bayésienne obtient de meilleurs résultats pour tous les tests. Selon le test des sites gagnants, l'approche bayésienne reconstruit de manière significativement différente 206 arbres, alors que la méthode du maximum de vraisemblance en reconstruit 233 arbres différents de l'arbre original. Pour le test des rangs signés de Wilcoxon et le test-t pour échantillons appariés, l'approche bayésienne construit différement 96 et 46 arbres respectivement, alors que la méthode du maximum de vraisemblance compte 125 et 48 arbres différents. Finalement le dernier test de matrice des distances, nous permet de constater que la distance entre les arbres reconstruits et l'arbre original est en moyenne de 12,718 pour l'approche bayésienne et de 47,727 pour la méthode par maximum de vraisemblance. Dans le contexte de la comparaison directe (voir la sous-section 5.3.1), l'approche bayésienne l'emporte 665 fois contre 335 pour la méthode du maximum de vraisemblance.

TABLEAU 5.3.3. Résultats de la deuxième simulation pour la méthode du maximum de vraisemblance (EMV) et l'approche bayésienne en fonction des tests de comparaison, à $\alpha = 5\%$

Méthode \ Test	Sites gagnants nombre d'arbres rejetés	Rangs signés de Wilcoxon nombre d'arbres rejetés	t nombre d'arbres rejetés	Matrice des distances distance moyenne
EMV	233	125	48	47,727
Bayes	206	96	46	12,718

5.3.3. Simulations 3, 4, 5 et 6

Dans cette sous-section, nous allons donner deux tableaux récapitulatifs pour chacune des simulations énumérées au début de la présente section. Les tableaux 5.3.4 et 5.3.5, nous permettent de constater que l'approche bayésienne reconstruit mieux les arbres que la méthode du maximum de vraisemblance. Pour les comparaisons directes pour chaque jeu de données reconstruit, nous obtenons des résultats qui indiquent également que l'approche bayésienne reconstruit de meilleurs arbres. Les résultats sont les suivants : pour la troisième simulation, nous obtenons 544 pour l'approche bayésienne contre 456 pour la méthode du maximum de vraisemblance; pour la quatrième (cinquième et sixième) simulation, l'approche bayésienne l'emporte 536 (625 et 696) contre 464 (375 et 304) pour la méthode du maximum de vraisemblance. Les deux approches semblent avoir plus de difficulté pour les simulations 3 et 4. Il est bien connu que la reconstruction d'arbres aux courtes branches entre les espèces est difficile à accomplir (voir Felsentein, 2004).

5.3.4. Test-t pour comparer les matrices des distances des deux méthodes

Pour vérifier qu'il existe bien une différence significative entre les arbres obtenus à l'aide des deux méthodes, nous construirons un test-t pour comparer les matrices des distances des deux méthodes. Les tableaux 5.3.6 à 5.3.8 montrent les différences entre les deux méthodes pour chacune des simulations. Comme il est

TABLEAU 5.3.4. Résultats des simulations pour la méthode du maximum de vraisemblance en fonction des tests de comparaison, à $\alpha = 5\%$

simulation \ Test	Sites gagnants nombre d'arbres rejetés	Rangs signés de Wilcoxon nombre d'arbres rejetés	t nombre d'arbres rejetés	Matrice des distances distance moyenne
1	335	159	35	30,220
2	233	125	48	47,727
3	826	554	53	29,210
4	698	527	62	39,752
5	412	161	49	118,400
6	322	121	57	132,947

TABLEAU 5.3.5. Résultats des simulations pour l'approche bayésienne en fonction des tests de comparaison, à $\alpha = 5\%$

simulation \ Test	Sites gagnants nombre d'arbres rejetés	Rangs signés de Wilcoxon nombre d'arbres rejetés	t nombre d'arbres rejetés	Matrice des distances distance moyenne
1	271	97	30	6,937
2	206	96	46	12,718
3	800	419	23	14,041
4	665	325	27	18,427
5	378	102	26	26,557
6	301	98	29	29,012

possible de le remarquer, pour toutes les simulations une différence significative est présente (à un niveau $\alpha = 5\%$). Donc, nous pouvons affirmer que les arbres construits par une approche bayésienne sont en moyenne toujours plus près de l'arbre original que ceux produits par la méthode du maximum de vraisemblance. De plus, nous pouvons également affirmer, en observant les écarts types, que les arbres obtenus par l'approche bayésienne sont moins variables (voir les tableaux 5.3.7 et 5.3.8).

TABLEAU 5.3.6. Résultats des test-t à échantillons indépendants pour comparer les matrices des distances obtenus par les deux méthodes en fonction des simulations (où $\bar{d}$ représente la distance moyenne et s_d l'écart type de la distance)

simulation	$\bar{d}$ EMV	s_d EMV	$\bar{d}$ Bayes	s_d Bayes	t	valeur-p
1	30,220	123,097	6,937	8,926	5,966	$< 0{,}0001$
2	47,727	165,688	12,718	11,678	6,665	$< 0{,}0001$
3	29,210	23,659	14,041	3,525	20,053	$< 0{,}0001$
4	39,752	32,714	18,427	5,129	20,365	$< 0{,}0001$
5	118,400	289,524	26,557	14,849	10,018	$< 0{,}0001$
6	132,947	250,126	29,012	17,751	13,107	$< 0{,}0001$

TABLEAU 5.3.7. Statistiques des matrices des distances obtenus par la méthode de maximum de vraisemblance en fonction des simulations (où $\bar{d}$ représente la distance moyenne et s_d l'écart type de la distance)

simulation	$\bar{d}$	s_d	médiane	minimum	maximum
1	30,220	123,097	5,95558	2.897	1740,789
2	47,727	165,688	11,265	0,872	1789,258
3	29,210	23,659	16,641	2,242	383,665
4	39,752	32,714	29,818	1,928	1576,887
5	118,400	289,524	95,682	3,374	1952,295
6	132,947	250,126	119,228	5,219	1053,958

TABLEAU 5.3.8. Statistiques des matrices des distances obtenus par l'approche bayésienne en fonction des simulations (où $\bar{d}$ représente la distance moyenne et s_d l'écart type de la distance)

simulation	$\bar{d}$	s_d	médiane	minimum	maximum
1	6,937	8,926	4,496181	1,757	44,237
2	12,718	11,678	9,098	1,738	44,678
3	14,041	3,525	15,060	1,932	20,151
4	18,427	5,129	35,225	0,935	36,786
5	26,557	14,849	27,872	1,292	72,159
6	29,012	17,751	39,117	0,892	85,206

5.4. Simulation d'un vrai jeu de données

Dans cette section du chapitre 5, nous allons appliquer les théories développées dans les quatre premiers chapitres à un vrai jeu de données. Nous allons encore comparer les deux méthodes, soit le maximum de vraisemblance et l'approche bayésienne, mais cette fois nous utiliserons un vrai jeu de données pour pouvoir les comparer. Ce vrai jeu de données sera composé de primates : soit l'humain (*Homo sapiens*), l'orang-outan (*Pongo pygmaeus abelii*), le gibbon (*Hylobates lar*), le gorille (*Gorilla gorilla*) et le chimpanzé (*Pan troglodytes*).

Définition 5.4.1. *D'une manière simpliste, le terme primate désigne les hommes, les singes et les lémuriens. Les primates sont formés de mammifères placentaires, caractérisés par une vie en général arboricole, des ongles aux doigts et orteils, la préhension par opposition du pouce, une prédominance de la vision sur l'olfaction. À ces évolutions s'ajoute chez l'homme le passage de la marche quadrupède à la bipédie.*

La phylogénie du groupe est bien établie en général (voir figure 5.4.1), mais le cas particulier de la position de l'espèce humaine a donné lieu à de nombreux débats. La classification traditionnelle réserve à l'homme la famille des hominidés, regroupant les deux espèces de chimpanzés, le gorille et l'orang-outan dans la famille des pongidés. Cette approche n'est plus retenue, les pongidés n'étant pas alors monophylétiques, car il semble bien établi que les hommes partagent avec les chimpanzés et le gorille un ancêtre distinct de ceux de l'orang-outan. Le groupe frère des humains a fait longtemps débat entre les chimpanzés, le gorille ou l'ensemble chimpanzés-gorille. Ce débat semble converger vers la première hypothèse (Schwartz, 1986).

5.4.1. Les données

Le choix des données s'avéra être une tâche ardue, puisque l'objectif de notre étude est de comparer la phylogénie que les biologistes ont trouvé à partir des caractères morphologiques et de la comparer à l'analyse phylogénétique faite à partir de données moléculaires. Dans le cadre de notre recherche, nous avons décidé d'étudier deux séquences de nucléotides mitochondriales, soit le 12S et le

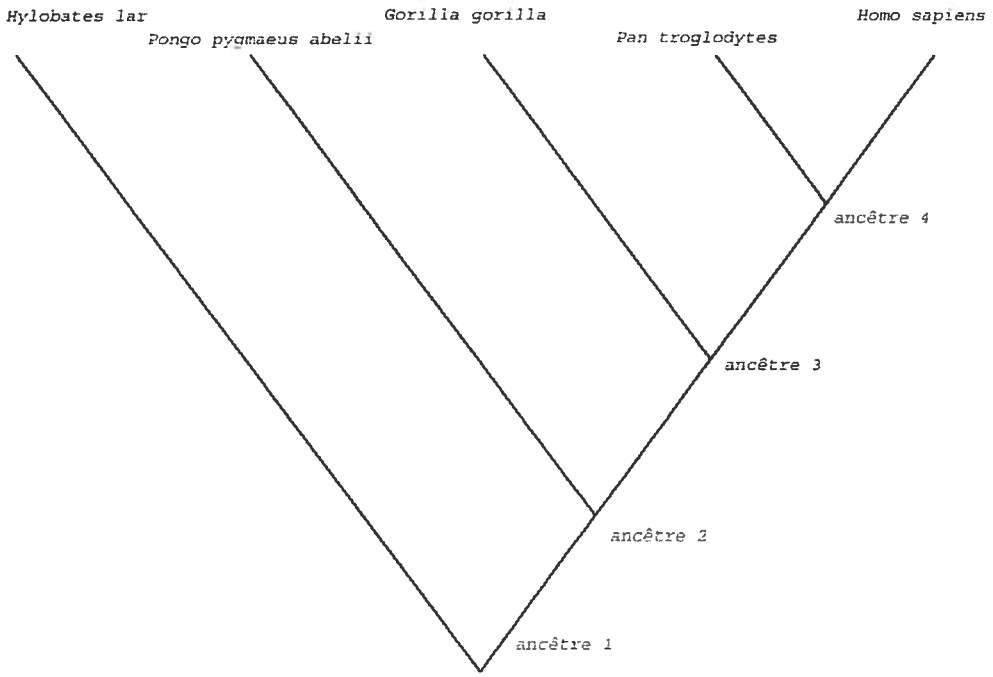


FIGURE 5.4.1. Phylogénie des primates (théorie contemporaine)

16S. Ces données furent obtenues à partir de la banque de données GenBank du National Center for Biotechnology.

Les nucléotides de nos deux gènes mitochondriaux seront nos caractères d'étude. L'état des caractères moléculaires sont : A (adénine), C (cytosine), G (guanine) et T (thymine). L'alignement des deux gènes a été exécuté sur ClustalX, un programme d'alignement de nucléotides (voir le site internet : http ://www.csc.fi/molbio/progs/clustalw/). Les paramètres qui ont servi à faire l'alignement sont les suivants : pairwise alignement = slow accurate, gap-opening =15, gap-extension=6,66, DNA transition weight = 0,5, delay divergent sequence (%) = 30. Après l'alignement, nous obtenons une base de données moléculaires de 2833 caractères.

5.5. RÉSULTATS D'UN VRAI JEU DE DONNÉES

Dans cette section, nous discuterons des résultats obtenus à partir de nos deux méthodes.

TABLEAU 5.5.1. Résultats de l'arbre phylogénétique des primates avec la méthode du maximum de vraisemblance (* = méthode plus exhaustive)

simulation \ statistique	Le logarithme de la vraisemblance	temps de simulation (approximatif)
JC	-6502,023	5 minutes
F81	-6450,279	6 minutes
K2P	-6395,593	6 minutes
JC*	-6400,681	1 heures 32 minutes
F81*	-6392,972	2 heure 9 minutes
K2P*	-6355,555	2 heure 35 minutes

5.5.1. Phylogénie obtenue à partir de la méthode du maximum de vraisemblance

Ainsi, en utilisant le modèle de Jukes-Cantor (JC, voir 3.2.1), le modèle de Felsenstein (F81, voir 3.2.3) et le modèle de Kimura à deux paramètres (K2P, voir 3.2.2), nous obtenons l'arbre illustré par la figure 5.5.1 (pour les trois méthodes l'arbre est le même) et les tableaux 5.5.1 et 5.5.2. Nous pouvons remarquer que le tableau 5.5.1 contient des résultats plus exhaustifs, représentés par le symbole *. Ces résultats sont obtenus en laissant fonctionner notre programme sans contrainte de temps (donc il visite plus d'arbres aléatoires). L'optimisation faite à partir de la routine *fminsearch* de Matlab prendra plus de temps à converger, mais donne des valeurs plus optimales. Ceci étant dit, les arbres obtenus sont les mêmes, seules les longueurs des branches varient pour donner un arbre plus optimal. En analysant l'arbre phylogénétique, nous observons que l'arbre obtenu est le même que celui imaginé par la théorie biologiste (voir la figure 5.4.1). Les espèces étant suffisament séparées par le temps, il est facile d'obtenir de tels résultats. Toutefois, il est intéressant de voir que notre programme fonctionne bien avec des vrais jeux de données. Cependant, il nous reste encore à appliquer ce jeu de données à l'approche bayésienne.

TABLEAU 5.5.2. L'arbre optimal obtenu par la méthode du maximum de vraisemblance (K2P*)

Branche	Espèces liées	longueur de la branche
1	Gibbon et ancêtre 1	0,082045
2	ancêtre 1 et ancêtre 2	0,014228
3	Orang-outan et ancêtre 2	0,064232
4	ancêtre 2 et ancêtre 3	0,022254
5	Gorille et ancêtre 3	0,034665
6	ancêtre 3 et ancêtre 4	0,006732
7	Chimpanzé et ancêtre 4	0,021227
8	Humain et ancêtre 4	0,028474

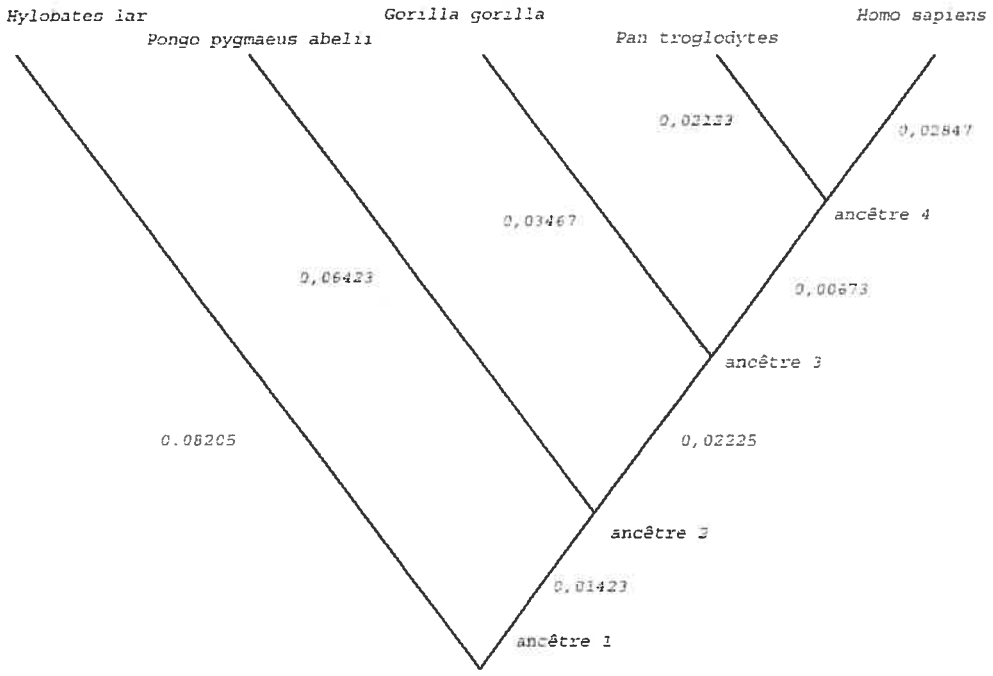


FIGURE 5.5.1. Phylogénie des primates (méthode du maximum de vraisemblance)

5.5.2. Phylogénie obtenue à partir de l'approche bayésienne

Contrairement à la sous-section précédente, nous allons maintenant trouver l'arbre phylogénétique à l'aide d'un processus bayésien (voir la section 4.2.2).

TABLEAU 5.5.3. Résultats de l'arbre phylogénétique des primates avec l'approche bayésienne

simulation \ statistique	Le logarithme de la vraisemblance	temps de simulation (approximatif)
JC	-6400,682	36 minutes
F81	-6352,126	1 heures 24 minutes
K2P	-6308,112	2 heures 3 minutes

Le processus est un peu plus long que la méthode de maximum de vraisemblance (procédé simple). Toutefois, en comparant les temps de simulation de la méthode de maximum de vraisemblance par un procédé plus exhaustif et ceux obtenus par l'approche bayésienne, nous remarquons que les temps de simulation de l'approche bayésienne sont moins longs. De plus, l'approche bayésienne nous donne beaucoup plus d'information sur notre arbre phylogénétique, puisqu'il échantillonne sur l'espace des arbres. Ainsi, nous obtenons un arbre avec un intervalle de crédibilité (voir arbre de consensus 4.2.4), ceci est représenté par un nombre entre parenthèses dans la figure 5.5.2; ceci donne la proportion de fois que deux espèces se regroupent dans un même noeud (dans l'algorithme Metropolis-Hastings). Les tableaux 5.5.3 et 5.5.4 donnent les arbres MAP obtenus par l'entremise de l'approche bayésienne. Les trois modèles utilisés donnent le même arbre phylogénétique (voir la figure 5.5.2), ils ne diffèrent que pour la longueur des branches. On peut remarquer que la vraisemblance est plus élevée pour l'arbre bayésien (soit un $\log L = -6308{,}1$ pour l'approche bayésienne et un $\log L = -6355{,}6$ pour la méthode du maximum de vraisemblance, K2P*). De plus, l'arbre consensus nous dit que 100 % des arbres sélectionnés par l'algorithme Metropolos-Hastings possédaient les mêmes noeuds que notre arbre MAP (il est important de noter que nous aurions obtenu le même résultat avec la méthode bootstrap). En ce qui concerne les longueurs des branches, nous pouvons dire qu'elles sont très similaires pour les deux méthodes.

TABLEAU 5.5.4. L'arbre optimal obtenu par l'approche bayésienne (K2P)

Branche	Espèces liées	longueur de la branche
1	Gibbon et ancêtre 1	0,081245
2	ancêtre 1 et ancêtre 2	0,012228
3	Orang-outan et ancêtre 2	0,064682
4	ancêtre 2 et ancêtre 3	0,022262
5	Gorille et ancêtre 3	0,035520
6	ancêtre 3 et ancêtre 4	0,006688
7	Chimpanzé et ancêtre 4	0,022169
8	Humain et ancêtre 4	0,028151

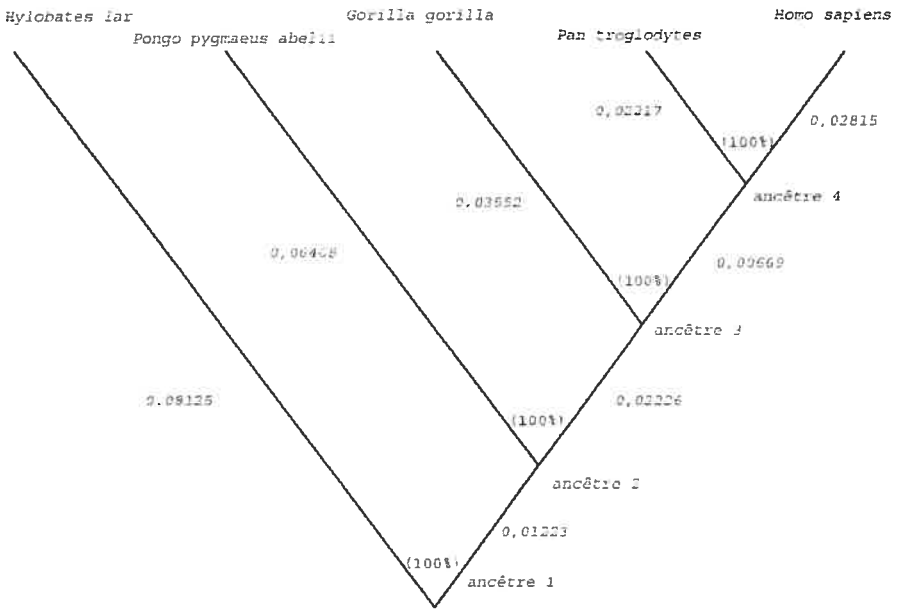


FIGURE 5.5.2. Phylogénie des primates obtenu avec l'approche bayésienne (les nombres entre parenthèses représentent le consensus du noeud)

5.6. CONCLUSION DU CHAPITRE

Dans ce dernier chapitre, nous avons appliqué les théories développées dans les quatre premiers chapitres. La méthode par maximum de vraisemblance et l'approche bayésienne furent comparées sous deux niveaux. Au premier niveau, nous

avons confronté les deux méthodes par l'entremise d'une reconstruction des jeux de données à partir des modèles utilisés. Par cette comparaison, il fut possible d'illustrer que l'approche bayésienne reconstruisait le jeu de données mieux dans tous les cas. Nous avons confronté les reconstructions à l'aide de plusieurs tests statistiques qui donnaient tous la faveur à l'approche bayésienne. En deuxième lieu, nous avons utilisé un vrai jeu de données pour comparer les deux méthodes. Ainsi, il a été possible de constater les différences et illustrer les avantages de l'approche bayésienne vis-à-vis la méthode du maximum de vraisemblance. Nous avons vu que les deux approches trouvaient le même arbre que les biologistes avaient imaginés. Toutefois, nous avons observé que l'arbre obtenu par l'approche bayésienne possédait une plus grande vraisemblance, donc une meilleure optimisation des branches. De plus, nous avons remarqué que l'approche bayésienne donnait une crédibilité à l'arbre, un consensus, qui nous permettait d'avoir une certaine validation sur l'arbre. Cette validation ne peut être obtenue par la méthode par maximum de vraisemblance, puisqu'elle visite les arbres de manières aléatoires. Un autre avantage de l'approche bayésienne est que cette dernière prend moins de temps à trouver un arbre phylogénétique. En conséquence, nous avons illustré à l'aide de données simulées et de vraies données que l'approche bayésienne est supérieure à la méthode du maximum de vraisemblance qui est considérée comme étant la norme en phylogénétique (la méthode la plus employée). Toutefois, il aurait été intéressant de comparer notre méthode de proposition d'arbres aux mécanismes (voir section 4.2.2.2) qui existent sur le marché (Mr.Bayes, 2005). Nous avons pu les comparer pour le vrai jeu de données, et les résultats sont les mêmes. Cependant, pour pouvoir les comparer de manière appropriée, nous aurions eu à reprogrammer leurs mécanismes de transition dans notre programme et refaire l'expérience des reconstructions des jeux de données. Ceci n'étant pas l'objectif du présent mémoire, nous laissons cette partie ouverte à des futurs chercheurs.

CONCLUSION

Le but de ce mémoire était de développer une approche bayésienne nous permettant de reconstruire un arbre phylogénétique à partir d'un jeu de données moléculaires. Dans un premier temps, nous avons expliqué les fondements de la phylogénétique et démontré les méthodes utilisées dans la reconstruction d'arbres phylogénétiques. Ensuite, nous avons élaboré nos procédés de représentation informatique des arbres et illustré notre méthode pour calculer la vraisemblance des arbres phylogénétiques. De plus, nous avons fait la démonstration de l'obtention des modèles de substitution et illustré nos méthodes d'optimisation des longueurs de branche par l'algorithme EM. Par la suite, nous avons décrit notre approche bayésienne et ses propriétés. Finalement, nous avons présenté les tests de comparaison employés qui nous permettent de valider notre approche bayésienne.

Les résultats obtenus au cinquième chapitre confirment que l'approche bayésienne possède d'énormes avantages vis-à-vis la méthode du maximum de vraisemblance. Nous avons vu que l'approche bayésienne est mathématiquement très similaire à celle du maximum de vraisemblance. Les équations sont très similaires et les modèles de substitution sont les mêmes. Les deux ne diffèrent que dans l'utilisation d'une distribution *a priori* sur les paramètres de l'arbre. L'approche bayésienne utilise la distribution *a priori* pour estimer la topologie des arbres et la longueur des branches. Donc, elle permet aux biologistes de placer plus d'importance sur certains arbres qui leurs semblent plus probables. Ceci permet de sauver énormément de temps de simulation, puisqu'en général les arbres les plus probables sont ceux qui intuitivement le sont. Nous avons vu que l'économie de temps est énorme. De plus, nous avons démontré que la vraisemblance de l'arbre MAP obtenue par l'approche bayésienne est supérieure à celle de l'arbre obtenue

par la méthode du maximum de vraisemblance. De plus, en échantillonnant sur des arbres indépendants, la méthode MCMC nous donne un arbre de consensus ce que la méthode du maximum de vraisemblance ne produit pas. Par l'entremise du test des sites gagnants, du test des rangs signés de Wilcoxon, du test t et de la matrice des distances, nous avons démontré que la reconstruction des jeux de données est de loin meilleure lorsque nous utilisons l'approche bayésienne.

En terminant, énumérons quelques pistes de recherche. Tout d'abord, l'approche bayésienne présentée dans le présent mémoire est construite à partir de notre méthode de proposition d'un nouvel arbre. Toutefois, nous aurions pu utiliser d'autres méthodes de proposition d'un nouvel arbre pour reconstruire les arbres phylogénétiques. Nous avons construit un programme informatique qui donne de bonnes solutions si nous les comparons aux autres programmes qui existent sur le marché, tels que PAUP, PHYLIP, MrBayes, etc. Cependant, notre programme comporte certaines contraintes de fonctionnement. La structure du programme nous empêchait d'avoir la possibilité de reconstruire un arbre phylogénétique pour 12 espèces et plus à la fois. Pour la reconstruction d'un arbre à six espèces le temps de simulation était raisonnable, mais à partir de sept espèces le temps de simulation s'étendait sur un grand nombre d'heures. Donc, il aurait été intéressant de remédier à ce problème de programmation pour obtenir un programme comparable à ceux qui sont présents sur le marché.

Enfin, l'utilisation des autres propriétés dans le domaine bayésien sont des pistes de recherche qui auraient pu être intéressantes à explorer. Par exemple, nous aurions pu donner un intervalle de crédibilité à la longueur des branches. Également, nous aurions pu créer une méthode qui lors du parcours de la chaîne de Markov optimise le choix du modèle de substitution. Faudrait-il utiliser une autre méthode de rééchantillonnage (que le MCMC) ou développer encore plus la théorie sur la phylogénétique bayesienne ? Ainsi, comme nous pouvons le remarquer l'approche bayésienne est une application récente dans le domaine de la phylogénétique. Celle-ci pourrait éventuellement être élargie de manière à pouvoir intégrer de nouveau procédé de sélection d'arbres phylogénétiques. Qui sait ? Elle pourrait devenir la nouvelle norme en phylogénétique.

Annexe A

A.1. Programmation

Cette annexe contient une partie des programmes utilisés pour la génération des résultats présentés dans ce mémoire. Précisons que le logiciel utilisé est *Matlab 7.0* (avec l'ajout du *Bioinformatics toolbox*).

A.1.1. Programmes

```
% Programme de depart

% Cette fonction cherche le fichier de donnees avec le programme donnees,
% elle le place dans une matrice ou chaque ligne est une espece avec ces
% genes dans chaque colonne

str = 'donnees.txt';
nbespV = 5;
nbgenes = 2833;

genMat = donnees(nbespV,nbgenes);

% Nous pouvons obtenir le maximum de vraisemblance avec le programmme Maxvraisemblance.
% Maxvraisemblance prend les arbres de la fonction aleatoire et calcule
% la fonction de vraisemblance pour chacun des arbres en
% permutatant les branches initiales. Elle prend le maximum parmi ces arbres
% et calcule un nombre aleatoire de permutations sur l'arbre maximum pour trouver
% l'arbre maximum

Maxvraisemblance(nbespV,nbgenes,genMat);

% Nous pouvons obtenir une methode bayesienne avec le programme Bayes. Le programme
% tente de creer un espace pour pouvoir se promener dans l'espace des arbres
% (reechantillonnage). Il commence par selectionner un nombre aleatoire de noeuds
% (avec le poids associe au nombre de noeuds - Donc, plus le nombre d'arbres pour
% un nombre de noeuds est eleve, plus de chance ce nombre de noeuds risque d'etre
% choisi). Ensuite, on fait des permutations pour evaluer la vraisemblance sur
% tous les arbres avec le nombre de noeuds choisi. Finalement, nous prennons l'arbre
% MAP et nous trouvons l'arbre de consensus a l'aide du MCMC

Bayes(nbespV,nbgenes,genMat);
```

```
function (mapMLE, mapBRANCHE, mapTEMPS)  = bayes(nbespV,nbgenes,genMat)


% MCMC :
% 1. Nous prennons un arbre aleatoire (Ti)
% 2. Ensuite, nous prennons un arbre voisin (Tj)
% 3. Calculer R = P(Tj)/P(Ti)
% 4.1 Si R >= 1, nous acceptons le nouvel arbre Tj
% 4.2 Si R < 1,  nous prennons un nombre x->U(0,1).
%     Si x < R, nous acceptons le nouvel arbre Tj
% 5. Sinon, nous rejettons Tj et nous conservons l'arbre Ti
% 6. Nous recommencons l'algorithme a partir de l'etape 2

[labelA,mleA,cataA,RotA,brancheA,tempsA] = voisin(nbespV,nbgenes,genMat);

          mapMLE = mleA;
          mapLABEL = labelA;
          mapCATA = cataA;
          mapROT = RotA;
          mapBRANCHE = brancheA;
          mapTEMPS = tempsA;

for iter=1:nbiter %nbiter : nombre d'iteration que nous desirons obtenir

      [labelN,mleN,cataN,RotN,brancheN,tempsN] = voisin(nbespV,nbgenes,genMat);

      if mleN < mleA
          mapMLE = mleN;
          mapLABEL = labelN;
          mapCATA = cataN;
          mapROT = RotN;
          mapBRANCHE = brancheN;
          mapTEMPS = tempsN;
      end

      % MCMC pour construire un arbre de consensus

     R = mleN/mleA;

     if  R >= 1
         mleA = mleN;
         labelA=labelN;
         cataA=cataN;
         RotA=RotN;
         brancheA = brancheN;
         tempsA = tempsN;
     end

     if  R < 1
         x = unifrnd(0,1);

         if x < R
             mleA = mleN;
             labelA=labelN;
             cataA=cataN;
             RotA=RotN;
             brancheA = brancheN;
             tempsA=tempsN;
```

```
        end
    end

    ecrireBranche(brancheA,fid5);

end
    fclose(fid5);


function  [label,mle,cata,Rot,brancheF,temps] = voisin(nbespV,nbgenes,genMat)

% algorithme en trois etapes pour selectionner l'arbre voisin
% 1. Construire une distribution a partir du poids associe a chaque noeud
%    (noeud -> U(0,1))
% 2. Faire une permutation des especes
% 3. Construire une distribution a partir des arbres du catalogue associe au
%    nombre de noeuds selectionne.
% La troisieme etape peut-être traduit de la maniere suivante:
% vous calculez la vraisemblance pour chacun des arbres avec le nombre de noeuds
% choisi et la permutation des especes initiales.

% 1. selection du nombre de noeuds

VectP = zeros(nbespV-1,1);

VectP = poidsNoeud(nbespV);
VectC = cumsum(VectP);

nbNoeud = size(VectP,2);

selpoids = unifrnd(0,1);
sel=0;

for i=1:nbespV-1
    if ((VectC(i) > selpoids) & (sel==0))
        noeud = i;
        sel = 1;
    end
end

% 2. Permutation des especes

[brancheF,strRot] = Rotation(branche,Rot,nbespV)

% 3. Poids des arbres par rapport au catalogue (reliee au nombre de noeuds)

[label,mle,cata,Rot,brancheF,temps]=lireFichier(nbespV,noeud,genMat);


function [labelF,mle,cataF,RotF,brancheF,temps]=lireFichier(nbespV,noeud,genMat)

% lireFichier :
% Une fonction entierement liee au deroulement de voisin.
% Elle va lire les arbres du catalogue pour le retransmettre a voisin.

    mle=0;
    label=1;

    str1 = ('abc', num2str(nbespV),'.txt');
```

```
    fid1 = fopen(str1,'r');

    Rot = randperm(nbespV);

    while 1

        tline = fgetl(fid1);

        if ~ischar(tline) , break, end
            getting = tline;

            nbligne = str2num((getting(1) getting(2)));
            nbcolonne = str2num((getting(3) getting(4)));

            compte = 5;
            cata = repmat(0,nbligne,nbcolonne);
            for m = 1:nbligne
                for n = 1:nbcolonne
                    cata(m,n) = str2num(getting(compte));
                    compte=compte+1;
                end
            end

            nbNoeudC = nbNoeudCat(cata,nbespV);

            if nbNoeudC == noeud
            strB='MLK';
            branche = Mat2Tree(cata);
            [brancheT,strRot] = Rotation(branche,Rot,nbespV);
            [mleT,tempsT]=ecrireNC(cata,brancheT,nbespV,noeud,genMat,label,strRot,strB);
                if mleT < mle
                    mle = mleT;
                    cataF = cata;
                    labelF = label;
                    RotF = Rot;
                    brancheF = brancheT;
                    temps = tempsT;
                end
            label = label + 1;
            end
    end


function [mle,temps]=ecrireNC(cata,brancheF,nbespV,nbNoeud,genMat,label,strRot,strB)

% ecrireNC :
% Une fonction entierement liee au deroulement de voisin.
% Elle va ecrire la vraisemblance de chaque arbre que lireFichier donne.

if strB == 'MLK'
    nbgenes = size(genMat,2)-1;
    [mle,temps] = vraisemblanceB(nbespV,nbgenes,brancheF,genMat);
end

    nbL = size(cata,1);
    nbC = size(cata,2);
    dim = nbL*nbC;
    temp = reshape(transpose(cata),1,dim);
```

```
    if strB == 'MLK'
     str2=('Catalogue', strB,' ',
           num2str(nbespV),'esp ',num2str(nbNoeud),'Rot',strRot,'.txt');
     fid2 = fopen(str2,'a');
    else
        str2 = ('Catalogue', strB,'.txt');
        fid2 = fopen(str2,'a');
    end

        fprintf(fid2,'%i',label);
        fprintf(fid2,'\t');

    if strB=='MLK'
        fprintf(fid2,'%e%',mle);
        fprintf(fid2,' \t');
    end

    if nbL >= 10
        fprintf(fid2,'%i',nbL);
    end

    if nbL < 10
        fprintf(fid2,'0%i',nbL);
    end

    if nbC >= 10
        fprintf(fid2,'%i',nbC);
    end

    if nbC < 10
        fprintf(fid2,'0%i',nbC);
    end


        fprintf(fid2,'%i',temp);
        fprintf(fid2,'\n');

        fclose(fid2);



function [vraisemblance,temps]=vraisemblance(nbespV,nbgenes,branche,genMat)

nbBranche = size(branche,1);
nbespA= max(branche(:,2))-nbespV;


% MATRICE DE TERMES
%   Cette fonction prend les branches pour ecrire les fonctions de chaque
%   caractere dans des matrices de termes.

%   nbBranche = nombre de branches de l'arbre
%   JC = Matrice Jukes Cantor
%   M = matrice des termes de l'equation

JC = repmat(0,2*nbBranche,1);

M = repmat(0,( 4^nbespA 2*nbBranche nbgenes ));
```

```
%                    les nucleotides
%                     en code ASCII
%
%
%                           A = 65
%                           C = 67
%                           G = 71
%                           T = 84

% ici VectNucl nous permet d'ecrire une matrice de toutes
% les possibilites pour les especes ancêtres
% VespBase = matrice des especes
% VespSite = matrice qui permet de joindre les matrices possibles
% associees aux especes ancestrales aux matrices des especes vraies.

VectNucl = (65 67 71 84);

VespBase=zeros(nbespV,nbgenes);
VespSite=zeros(nbespV, 4^nbespA );


for i=1:nbespV
    for j=1:nbgenes
        VespBase(i,j) = genMat(i,j+1);
    end
end



% Algorithme de construction de la matrice M
% permettant de differencier les especes
% VespSiteCom = melange especes Vraies et ancestrales tous les possibilites
% d'histoires sont presentees dans cette matrice


for i=1:nbgenes
    VespSite=repmat(VespBase(:,i),1, 4^nbespA );
    MatAnc = ArrRep(VectNucl,nbespA);
    VespSiteCom = (VespSite ; MatAnc);

    for k=1: 4^nbespA
           for j=1:nbBranche


            esp1 = branche(j,1);
            esp2 = branche(j,2);

            nuc1=VespSiteCom(esp1,k);
            nuc2=VespSiteCom(esp2,k);

            indB=j*2;

            if nuc1==nuc2
                M(k,indB,i)=1;
            end

            if nuc1~=nuc2
                M(k,indB-1,i)=1;
```

```
            end

        end
    end
end

% Algorithme permettant de trouver
% le maximum de vraisemblance de l'arbre
% voir fonction PF

    t = ();
    LT = ();
    ProbT = ();
    x = ();
    fval = ();

    a=0.5;
    b=0.5;

for j = 1:20

    t((1:nbBranche))=0.01+0.1*(j-1);

    for i=1:nbBranche
         JC(2*i-1,1)= log(0.25 - 0.25*exp(- t(i)));
         JC(2*i,1)= log(0.25 + 0.75*exp(- t(i)));
    end

    [x,fval]=fminsearch(@PFB,t,
             optimset('MaxIter',500,'MaxFunEvals',500),M,JC,nbBranche,nbgenes,a,b);

   LT(j,:)=x;
   ProbT(j,:)=fval;

end

ProbT;

[valeur, ind]=min(ProbT);
temps = LT(ind,:);
vraisemblance = valeur


function PFB = PFB(t,M,JC,nbBranche,nbgenes,a,b)

% Fonction de vraisemblance ecrite en langage Matlab

for i=1:nbBranche
    JC(2*i-1,1)= log(0.25 - 0.25*exp(- t(i)));
    JC(2*i,1)= log(0.25 + 0.75*exp(- t(i)));
end

TF = 1;

for i=1:nbgenes
    TF = TF*(sum(exp(M(:,:,i)*JC)));
end

PFB=(-1*TF+exp(-1000*min(sign((t)))))*((prod(exp(t*(a-1))))/(prod(1+exp(-t))^(a+b)));
```

```
function reconstruction = reconstruction(nbespV,nbgenes,branche,t)

% Cette fonction permet de reconstruire les jeux de donnees
% de comparer la difference entre la reconstruction bayesienne
% et celle obtenue par maximum de vraisemblance.

nbgenes = 2833;
nbespV = 5;

% arbre de depart (celui que nous voulons reconstruire)
branche= (1,6;2,7;3,8;4,9;5,10;6,7;7,8;8,9;9,10);
t = (1,1,1,1,1,1,1,1,1);

nbBranche = size(branche,1);
nbAnc = max(branche(:,2)) - nbespV;
VectNucl = (65 67 71 84);

replicate = 0;

for iter=1:1000
genMatA = repMat(0,nbespV+nbAnc,nbgenes);
genMatL = 0;

for i = 1:nbgenes
    %selection aleatoire du nucleotide de l'ancetre
    alea = unidrnd(4);
    NucAnc = VectNucl(alea);
    genMatA(nbespV+nbAnc,i)=NucAnc;


 for k=1:(nbespV+nbAnc)

    for j=1:nbBranche

        if branche(j,2)==(nbespV+nbAnc-k+1)
             espF = branche(j,2);
             espT = branche(j,1);

             nucl = genMatA(espF,i);

                ProbP = (0.25 + 0.75*exp(- t(j)));
                ProbC = 3*(0.25 - 0.25*exp(- t(j)));
                ProbT = ProbP + ProbC;
                selPC = unifrnd(0,1);

                if selPC <= ProbC
                    compte = 1;
                    for ii=1:4
                        if VectNucl(ii) ~= nucl
                            VectT(compte) = VectNucl(ii);
                            compte=compte+1;
                        end
                    end
                    NouvelNuc = VectT(unidrnd(3));
                    genMatA(espT,i)= NouvelNuc;
                end
```

```
                if selPC > ProbC
                    genMatA(espT,i)=nucl;
                end

            end
        end
    end
end

vectE = reshape((1:nbespV),nbespV,1);

genMatA
genMatF = (vectE genMatA(1:nbespV,:))

genMatL = transNaC(genMatF);

(mapMLE, mapBRANCHE, mapTEMPS)  = bayes(nbespV,nbgenes,genMatL)
branche
(maxMLE, mleBRANCHE, mleTEMPS) = Maxvraisemblance(nbespV,nbgenes,genMatL)

replicate = replicate + 1;

[pvalueT(iter),pvalueC(iter),pvalueWC(iter),distBL(iter)]
    = diffBETtree(branche,mapBRANCHE,t,mapTEMPS,genMatL,nbespV,nbgenes)

[pvalueTml(iter),pvalueCml(iter),pvalueWCml(iter),distBLml(iter)]
    = diffBETtree(branche,mleBRANCHE,t,mleTEMPS,genMatL,nbespV,nbgenes)

fid2 = fopen('sim2.txt','a')

       if replicate==1
       fprintf(fid2,'replicate, type, pvalueT, pvalueC, pvalueWC  \n')
       end

       fprintf(fid2,'%i, BAYES, %f, %f, %f, %f \n ',
    replicate, pvalueT(iter),pvalueC(iter),pvalueWC(iter),distBL(iter))

 fprintf(fid2,'%i, MLE  , %f, %f, %f, %f  \n ',
       replicate, pvalueTml(iter),pvalueCml(iter),pvalueWCml(iter),distBLml(iter))


    fclose(fid2)




function [pvalueT,pvalueC,pvalueWC,distBL]
          =diffBETtree(branche1,branche2,t1,t2,genMat,nbespV,nbgenes)

% Cette fonction permet de faire les quatre tests
% de comparaison vus au chapitre 4.

for iter=1:2

    if iter == 1
        branche = branche1;
        t = t1;
    end
```

```
    if iter == 2
        branche = branche2;
        t = t2;
    end

        nbBranche = size(branche,1);
        nbAnc = max(branche(:,2))-nbespV;
        VespBase = genMat(:,2:(nbgenes+1));
        VectN = (65 67 71 84);
        MatAnc = ArrRep(VectN,nbAnc);
        SommeS =0;

for i = 1:nbgenes

    VespSite=repmat(VespBase(:,i),1, 4^nbAnc );
    VespSiteCom = (VespSite ; MatAnc);
    SommeB = 0;
    SommeBT = 0;

    for k = 1:(4 ^nbAnc )
        for j=1:nbBranche

            esp1 = branche(j,1);
            esp2 = branche(j,2);

            nuc1=VespSiteCom(esp1,k);
            nuc2=VespSiteCom(esp2,k);

            if nuc1 == nuc2
                SommeB(j) = 0.25 + 0.75*exp(- t(j));
            end

            if nuc1 ~= nuc2
                SommeB(j) = 0.25 - 0.25*exp(- t(j));
            end

        end
       SommeBT(k) = prod(SommeB);

      end

      if  iter == 1
        SommeS1(i) = sum(SommeBT);
      end

      if  iter == 2
        SommeS2(i) = sum(SommeBT);
      end
  end

  if iter == 1
      lnL1 = log(SommeS1);
  end

  if iter == 2
      lnL2 = log(SommeS2);
  end
end
```

```
% test-t

d = lnL1 - lnL2;

dbar = mean(d);
EcartD = std(d);

ttest = dbar/(EcartD/(nbgenes ^0.5 ))

pvalueT = 1-tcdf(abs(ttest),nbgenes-1)

if pvalueT < 0.05
    'il y a une difference significative entre les 2 arbres selon le t-test'
else
    'il n y pas de difference significative entre les 2 arbres selon le t-test'
end

% Winning sites test

Win1 = 0;
Win2 = 0;
for i=1:nbgenes
    if lnL1(i)>lnL2(i)
        Win1 = Win1+1;
    else
        Win2 = Win2+1;
    end
end

(espB,varB) = binostat(nbgenes,0.5);

Chi2 = ((Win1-espB)^2 )/varB

pvalueC = 1-chi2cdf(Chi2,1)


if pvalueC < 0.05
    'il y a une difference significative entre les
     2 arbres selon le winning test (chi-carre)'
else
    'il n y pas de difference significative entre les
     2 arbres selon le winning test (chi-carre)'
end


% Wilcoxon signed ranks test

sommerang=0;
(B,ind) = sort(abs(d));

for i=1:nbgenes
    if d(ind(i))<0
        sommerang = i + sommerang;
    end
end

moyWC = (1/4)*nbgenes*(nbgenes+1);
```

```
ecartWC = ((nbgenes*(nbgenes+1)*((2*nbgenes)+1))/24)^0.5;

WCN = (sommerang-moyWC)/ecartWC

pvalueWC = 1-normcdf(abs(WCN),0,1)

if pvalueWC < 0.05
    'il y a une difference significative entre les
     2 arbres selon le Wilcoxon signed ranked test (normal app)'
else
    'il n y pas de difference significative entre les
     2 arbres selon le Wilcoxon signed ranked test (normal app)'
end


% matrice des distances entre especes

for iter=1:2

    if iter==1
        branche = branche1;
        t = t1;
    end

    if iter==2
        branche = branche2;
        t = t2;
    end

nbBranche = size(branche,1);
nbTot = max(branche(:,2));
nbAnc = nbTot - nbespV;

for i=1:nbespV
    espE = i;
    espS = 0;
    compte=1;
    while espS~=nbTot
        Trace2(i,compte)=espE;
        for j=1:nbBranche
            if branche(j,1)==espE
                espS = branche(j,2);
                indexBr(i,compte) = j;
            end
        end
        espE = espS;
        compte=compte+1;
        Trace2(i,compte)=espS;
    end
end

Trajet = repmat(0,nbespV,nbespV);
nbC = factorial(nbespV-1);
compteC = 1;
for i=1:(nbespV-1)
    nbTrace1 = size(Trace2(i,:),2);
    for j=i+1:nbespV
        nbTrace2 = size(Trace2(j,:),2);
        chemin=1;
```

```
        k=1;
        m=1;
        tm=1;
        tk=1;
            while Trace2(i,k)~=Trace2(j,m)
                k=k+1;
                m=1;
                while ((Trace2(i,k)~=Trace2(j,m)) & (m<nbTrace2))
                    m = m+1;

                    if Trace2(i,k)==Trace2(j,m)

                            tk=k;

                            while tk~=1
                                if iter==1
                                    Trajet1(compteC,chemin) = indexBr(i,tk-1);
                                end

                                if iter==2
                                    Trajet2(compteC,chemin) = indexBr(i,tk-1);
                                end

                                    chemin=chemin+1;
                                    tk = tk-1;
                            end

                            tm=m;

                            while tm~=1
                                if iter==1
                                    Trajet1(compteC,chemin) = indexBr(j,tm-1);
                                end

                                if iter==2
                                    Trajet2(compteC,chemin) = indexBr(j,tm-1);
                                end

                                    chemin=chemin+1;
                                    tm = tm-1;
                            end

                    end
                end
            end
        compteC=compteC+1;
    end
end

if iter==1
    Trajet=Trajet1;
end

if iter==2
    Trajet=Trajet2;
end

maxTrajet = size(Trajet,2);
compteC = 1;
```

```
for i=1:(nbespV-1)
    for j=i+1:nbespV
        sommeDist = 0;
       for k=1:maxTrajet
           if Trajet(compteC,k)~=0
               sommeDist = t(Trajet(compteC,k)) + sommeDist;
           end
       end

       if iter==1
        DistAO1(i,j) = sommeDist;
        DistAO1(j,i) = sommeDist;
       end

       if iter==2
        DistAO2(i,j) = sommeDist;
        DistAO2(j,i) = sommeDist;
       end

       compteC = compteC+1;
    end
end

end

MATdistBL = abs(DistAO1 - DistAO2);
distBL = sum(sum(MATdistBL)/2);
```

BIBLIOGRAPHIE

AGAPOW, P.M. (2003), *Introduction to Bayesian phylogenetics*, présenté au Royal Botanical Gardens Kew.

BATESON, W. (1902), *Mendel's Principles of Heredity*, Cambridge University Press, Londres.

BAYES, T. (1763), An essay towards solving a problem in the doctrine of chances, *Philosophical Transactions of the Royal Society*, 330-418.

CAMIN, J.H. ET SOKAL, R.R. (1965), A method for deducing branching sequences in phylogeny, *Evolution*, **19**, 311-326.

CAMPBELL, N. ET MATHIEU, R. (1995), *Biologie*, Renouveau pédagogique, Saint-Laurent.

DARLU, P. ET TASSY, P. (1993), *Reconstruction phylogénétique : concepts et méthodes*, Masson, Paris.

DARWIN, C. (1859), *The Origin of Species*, John-Murray, Londres.

DEMPSTER, A.P., LAIRD, N.M. ET RUBIN, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, **B 39**, 1-38.

DESTREMPES, F., MIGNOTTE, M. ET ANGERS, J.F. (2005), A stochastic method for Bayesian estimation of hidden markov field models with application to a color model, *IEEE Transactions on Image Processing*, **14(8)**, 1096-1124.

EDWARDS, A.W.F. ET CAVALLI-SFORZA, L.L. (1963), The reconstruction of evolution, *Annals of Human Genetics* , **27**, 105-106.

EDWARDS, A.W.F. ET CAVALLI-SFORZA, L.L. (1964), Reconstruction of evolutionary trees, *Phenetic and Phylogenetic Classification*, V.H. Heywood and J.McNeill Systematics Association Publication, Londres, 67-76.

KIMURA, M. (1981), Estimation of evolutionary distances between homologous nucleotide sequences, *Proceedings of the National Academy of Sciences*, **78**, 454-458.

LANAVE, C., PREPARATA, G., SACCONE, C. ET SERIO. G. (1984), A new method for calculating evolutionary substitution rates, *Journal of Molecular Evolution*, **20**, 86-93.

LARGET, B. SIMON, D.L. (1999), Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees, *Molecular Biology and Evolution*, **16**, 750-759.

LARGET, B., SIMON, D.L., KADANE, J.B. ET SWEET, D. (2005), A Bayesian analysis of metazoan mitochondrial genome arrangements, *Molecular Biology and Evolution*, **22(3)**, 486-495.

LI, S., PEARL, D.K. ET DOSS, H. (2000), Phylogenetic tree construction using Markov Chain Monte Carlo, *Journal of the American Statistical Association*, **95**, 493-508.

MAU, B. ET NEWTON, M.A. (1997), Phylogenetic inference for binary data on dendograms using Markov Chain Monte Carlo, *Journal of Computational and Graphical Statistics*, **6**, 122-131.

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H. ET TELLER, E. (1953), Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087-1092.

NEYMAN, J. (1971), Molecular studies of evolution : A source of novel statistical problems, *Statistical Decision Theory and Related Topics*, Academic Press, New York, 1-27.

RICE, J.A. (1995), *Mathematical Statistics and Data Analysis*, 2^e édition, Duxbury Press, Belmont.

ROBERT, C.P. (2001), *The Bayesian Choice*, 2^e édition, Springer-Verlag, New-York.

RODRIGUEZ, F., OLIVER, J.L., MARIN, A. ET MEDINA, J.R. (1990), The general stochastic model of nucleotide substitution, *Journal of Theoretical Biology*, **142**, 485-501.

SCHWARTZ, J. (1986), Primate systematics and a classication of the order in Comparative Primate Biology, *Systematics, Evolution and Anatomy*, **1**, 1-42.

SWOFFORD, D.L., THORNE, J.L. ET FELSENSTEIN, J. (1996a), The topology-dependent permutation test for monophyly does not test for monophyly, *Systematic Biology*, **45**, 575-579.

SWOFFORD, D.L., THORNE, J.L., FELSENSTEIN, J. ET HILLIS, D.M. (1996b), Phylogenetic inference, *Molecular Systematics*, David M. Hillis(éditeur) et Craig Moritz (éditeur), Sinauer Associates, Sunderland, pages 407-514.

TAMURA, K. ET NEI, M. (1993), Estimation of the number of nucleotide substitutions in the control region of mithondrial DNA in humans and chimpanzees, *Molecular Biology and Evolution*, **15**, 512-526.

TAVARÉ, S. (1986), Some probabilistic and statistical problems in the analysis of DNA sequences, *Lectures on Mathematics in the Life Sciences*, **17**, 57-86.

TEMPLETON, A.R. (1983), Phylogenetic inference from restriction endonuclease clevage site maps with particular reference to the evolution of humans and the apes, *Evolution*, **37**, 221-244.

WILSON, S. (1988), Measuring inconsistency in phylogenetics trees, *Journal of Theoretical Biology*, **190**, 15-36.

YANG, Z. ET RANNALA, B. (1997), Bayesian phylogenetic inference using DNA sequences : A Markov Chain Monte Carlo method, *Molecular Biology and Evolution*, **14**, 714-724.

ZHARKIKH, A. (1994), Estimation of evolutionary distances between nucleotide sequences, *Journal of Molecular Evolution*, **15**, 512-526.